



Enginius Segmentation

Khush Domadiya, California State University Eastbay

Copyright (c) 2025, DecisionPro Inc.

Table of contents

Warnings

Segmentation options

- Options selected
- Data description

Data transformation

Segment solution

- 2-segment solution
- Dendrogram
- Scree plot

Segment description

- Segment size
- Segment description
- Segmentation space
- Segment membership

Segment profiles

- Spider chart
- Segment 1 profile
- Segment 2 profile

Descriptor analysis

- Descriptors
- Descriptor space

Classification model

- Introduction
- Model coefficients
- P-values
- Confusion matrix
- Model predictions

Warnings

The following warnings were triggered during execution. Although they did not interrupt the analyses, they might indicate that there is an issue with the data or with the options chosen. Please review them carefully before going any further.

These variables from discriminant data are removed because of high collinearity with variables present in the report - Q6 = Iced coffee drinks, Hot coffee drinks, Q6 = Blended drinks (Freeze, Frost etc.), Q6 = Iced coffee drinks, Energy drinks (Rebel line), Tea / Lemonade / Other non-coffee drinks, Q6 = Blended drinks (Freeze,

Frost etc.), Tea / Lemonade / Other non-coffee drinks, Q6 = Hot coffee drinks, Tea / Lemonade / Other non-coffee drinks, Q6 = Iced coffee drinks, Hot coffee drinks, Tea / Lemonade / Other non-coffee drinks, Q6 = Iced coffee drinks, Blended drinks (Freeze, Frost etc.), Q6 = Blended drinks (Freeze, Frost etc.), Energy drinks (Rebel line), Tea / Lemonade / Other non-coffee drinks, Q6 = Iced coffee drinks, Energy drinks (Rebel line), Q6 = Iced coffee drinks, Blended drinks (Freeze, Frost etc.), Tea / Lemonade / Other non-coffee drinks, Q8 = Walk-up window, Q8 = Mobile app order, Q15 = 35-44 years old, Q15 = 45-54 years old, Q15 = Under 18, Q16 = Male, Q23 = Adventurous, Social, Creative, Q23 = Health-conscious, Adventurous, Creative, Q23 = Social, Busy, Creative, Q23 = Social, Relaxed, Q23 = Health-conscious, Relaxed, Trend-seeker, Q23 = Adventurous, Busy, Relaxed, Q23 = Health-conscious, Social, Busy, Q23 = Adventurous, Social, Relaxed, Q23 = Health-conscious, Busy, Trend-seeker, Q23 = Relaxed, Q23 = Health-conscious, Busy, Q23 = Busy, Creative, Relaxed, Q23 = Health-conscious, Busy, Relaxed

Segmentation options

Options selected

Option	Selection
Clustering method	Hierarchical
Standardization method	none
Segments forced	No
Run discriminant analysis	Yes
Run classification analysis	Yes
Date and time	2025-11-20 02:48:00 UTC

Options selected.

Data description

Data	Number of Rows	Number of columns	Column names
1 Segmentation data	17	7	\, Q9_1, Q9_2, Q9_3, Q9_4, ...
2 Discriminant data	17	9	\, Q6, Q8, Q15, Q16, ...

Data description.

Data transformation

Standardization has not been performed.

Segment solution

2-segment solution

The ideal number of segments is a function of statistical fit (what the data say), managerial relevance (what makes the most sense from a managerial point of view), and targetability (can the segments be easily targeted).

When the three criteria do not perfectly converge, selecting the right number of segments becomes a judgment call.

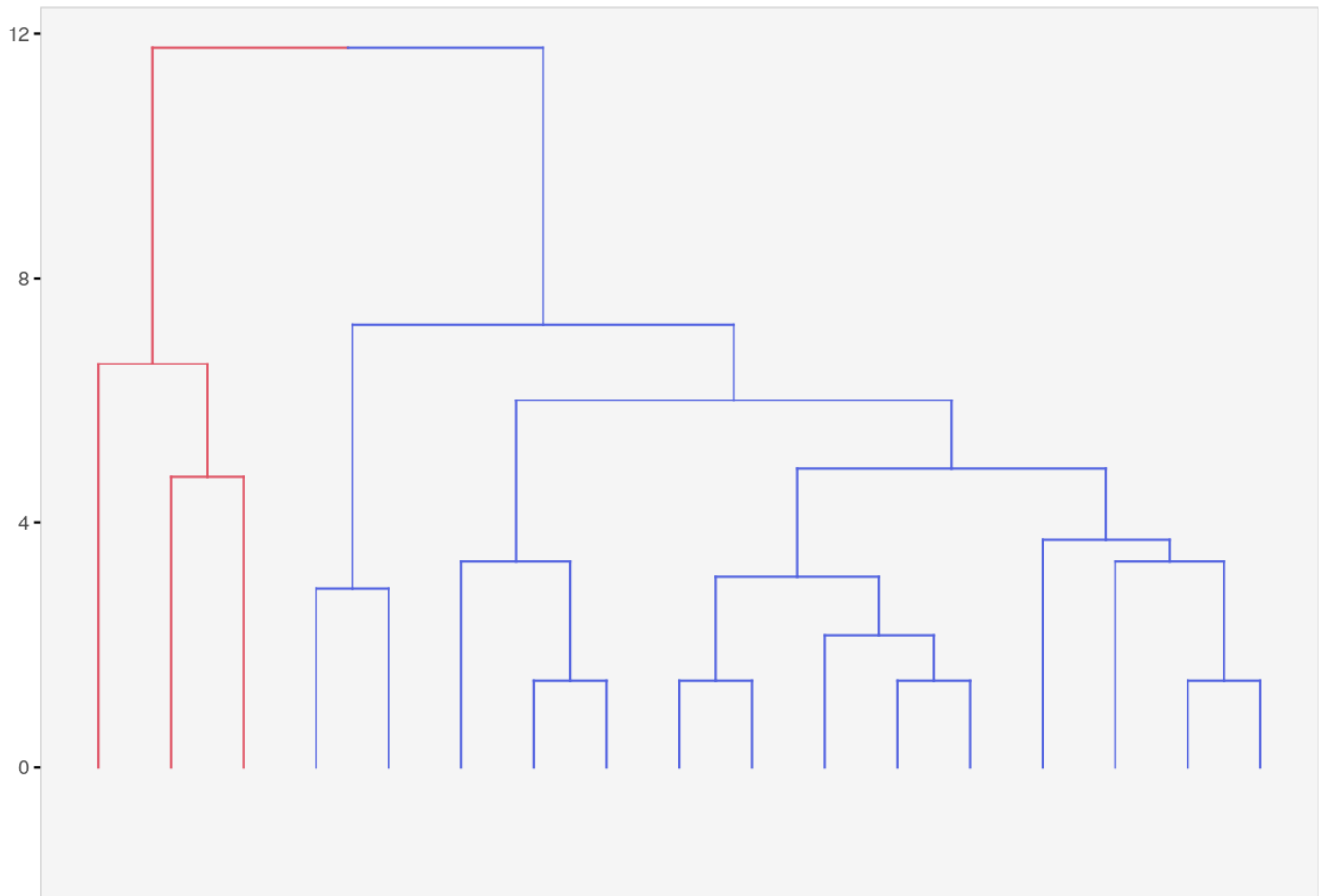
Using a statistical criteria exclusively (see scree plot analysis below), we have retained 2 segments.

The segmentation method relies on the hierarchical clustering approach. This approach generates a dendrogram that we display next.

Dendrogram

The dendrogram represents the grouping process of observations into clusters. The chart reads from bottom (all initial observations are separated) to top (all observations are clustered into one unique segment).

The height represents the distance between the two groups of observations being merged at each step. If two very distant groups are being merged, this will create a 'jump' in the dendrogram, indicating that it might be wise to stop the clustering process before.



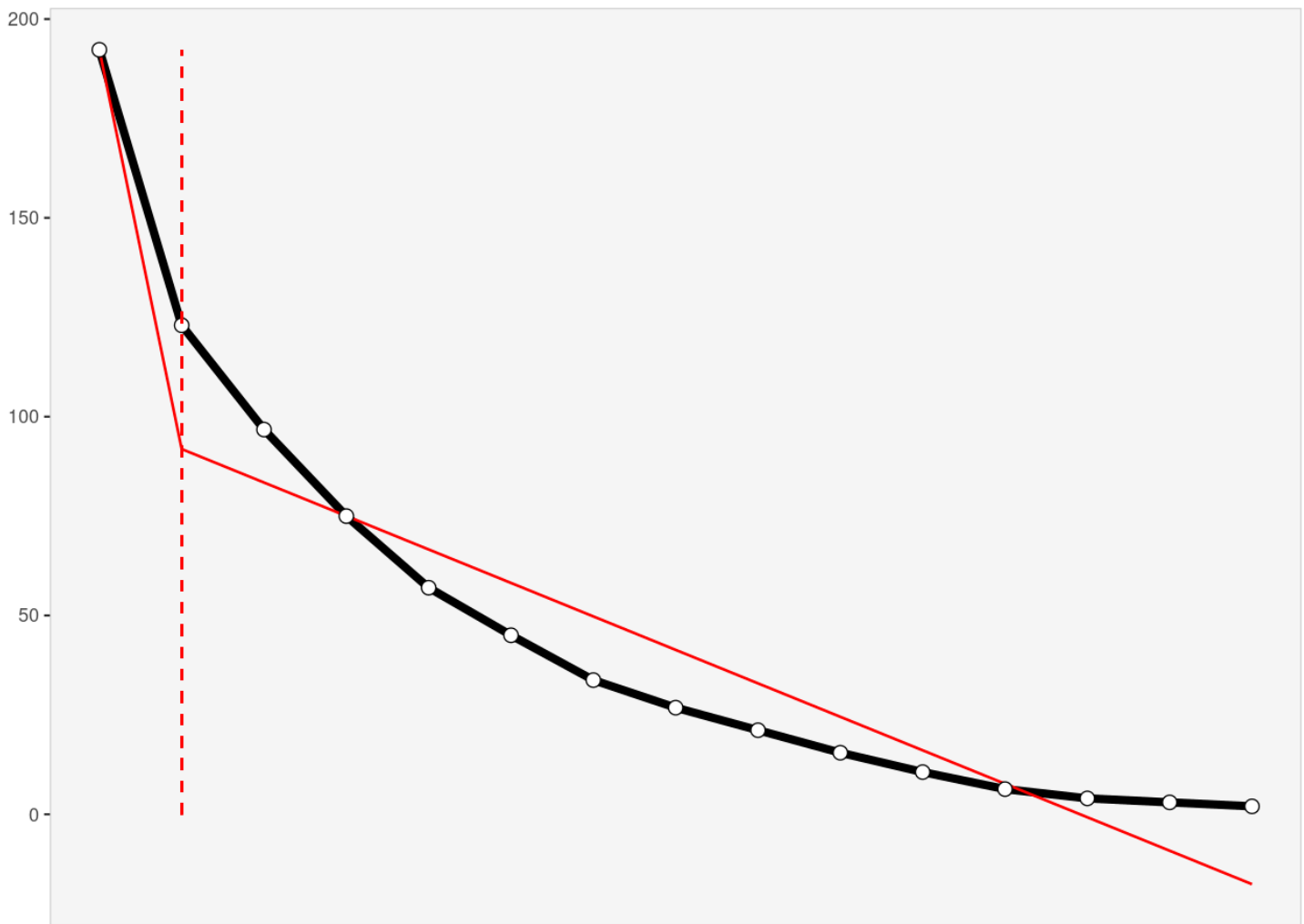
Dendrogram. The dendrogram is a tree diagram to illustrate the arrangement of clusters produced by hierarchical clustering, and how the observations are incrementally clustered together.

Scree plot

The screeplot displays, for each cluster solution, a measure of within-cluster heterogeneity. If clusters group observations that are widely different (which will happen if the number of clusters is too small to capture the variability in the data), the value will be high.

A good cluster solution might be where the screeplot displays an 'elbow', that is, where increasing the number of clusters beyond a certain point does not dramatically decreases within-cluster heterogeneity.

The measure displayed in the screeplot is related, but not equivalent, to the distance reported in the dendrogram.



Scree plot. The scree plot compares the sum of squared error (SSE) for each cluster solution. A good cluster solution might be when the SSE slows dramatically, creating an 'elbow'. Such elbow does not always exist. If number of segments is equal to maximum possible segments elbow cannot be created.

From a statistical point of view, the SSE reported in the screeplot is computed as the sum of squared error between each observation and its cluster centroid (or center), summed over all the observations.

Segment description

Segment size

	Population	Segment 1	Segment 2
Size	17	3	14
Relative size	100%	18%	82%

Segment size.

Segment description

	Population	Segment 1	Segment 2
Q9_1	3.8	0.333	4.543
Q9_2	4.29	2.67	4.64
Q9_3	4.18	4.33	4.14
Q9_4	3.47	3	3.57
Q9_5	3.17	2.42	3.33
Q9_6	3.88	2	4.29

Segment description. Average value of each segmentation variable, overall for each segment (centroid). Segmentation variables that are statistically different from the rest of the population are highlighted in red (lower) or green (higher).



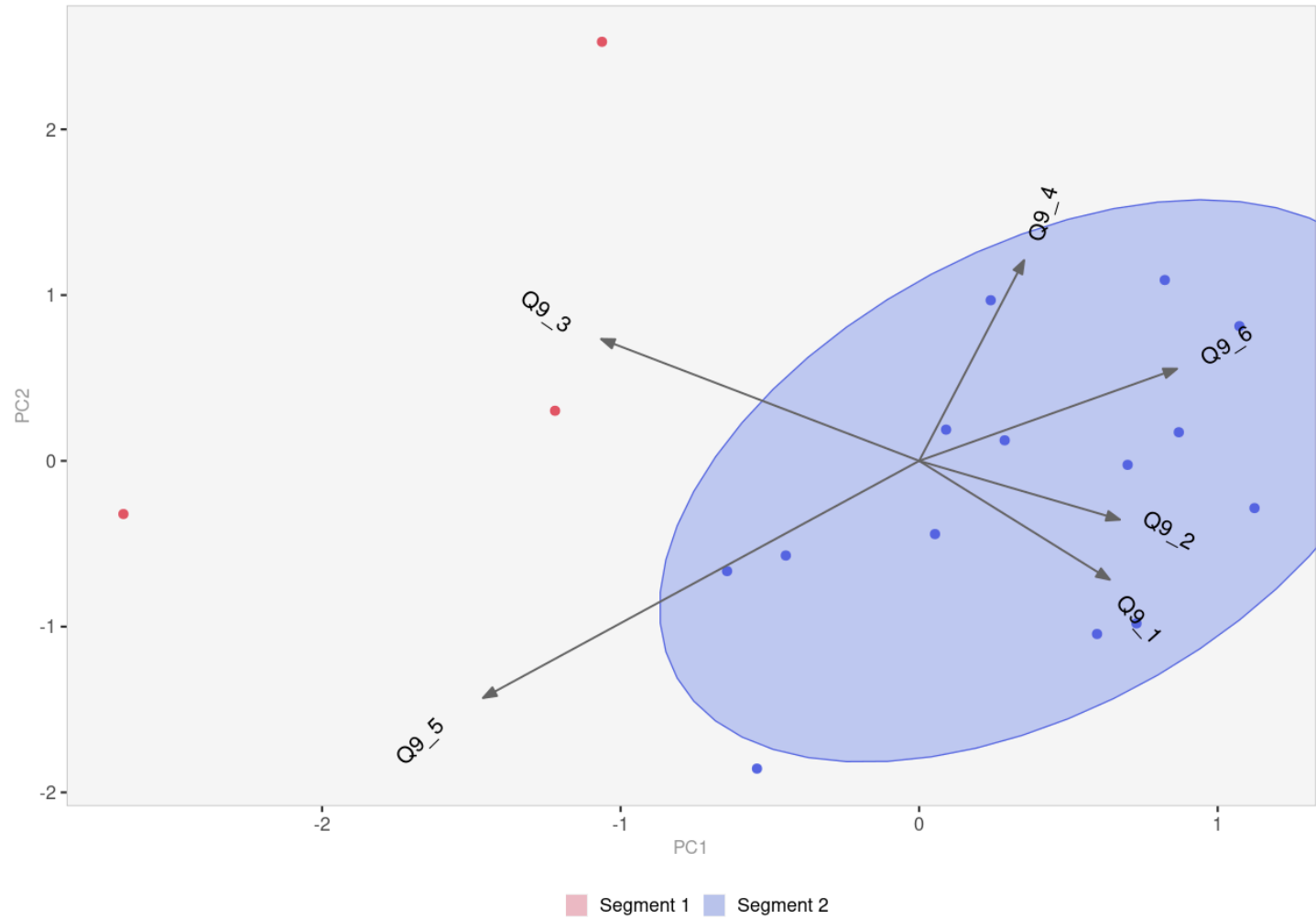
Segment differences per segment. Cell colors indicate to what extent a segment is statistically different from the rest of the population on each segmentation variable.

Segmentation space

The chart below is a graphical representation of the various segments, segment members, and segmentation variables. It is obtained by plotting the first two dimensions of a principal component analysis performed on the (standardized) segmentation data, on top of which segment information has been overlaid.

Because only the first two dimensions of the PCA are displayed, and these two dimensions capture only 63.8% of the variance in the data, some differences between segments might not appear here. Note that segmentation variables with no variance, if any, have been excluded.

Two clusters that appear to overlap on the first two dimensions might be distinct on other dimensions. Consequently, this chart is a useful guide, for checking which variables are correlated, but may be misleading if used to select the optimal number of segments.



Segment space. Spatial representation of segments and segmentation variables, using principal component analysis.

Segment membership

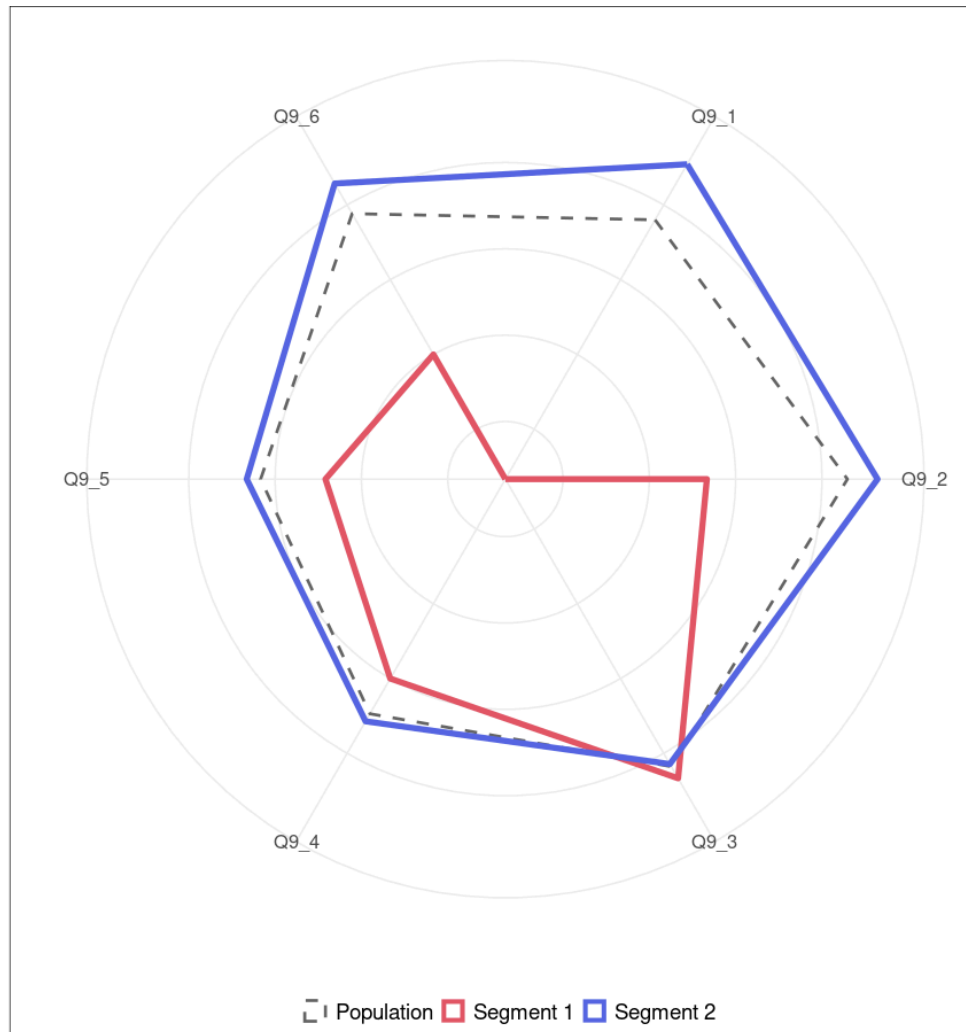
Segment	
1	1
2	2
3	2
4	2
5	2
6	2
7	2
8	2
10	1
11	2

Segment membership (excerpt). Segment to which each member of the population belongs to. The complete membership list is only available in the Excel formatted output.

Segment profiles

Spider chart

Spider chart comparing the averages of the segmentation variables across all segments.

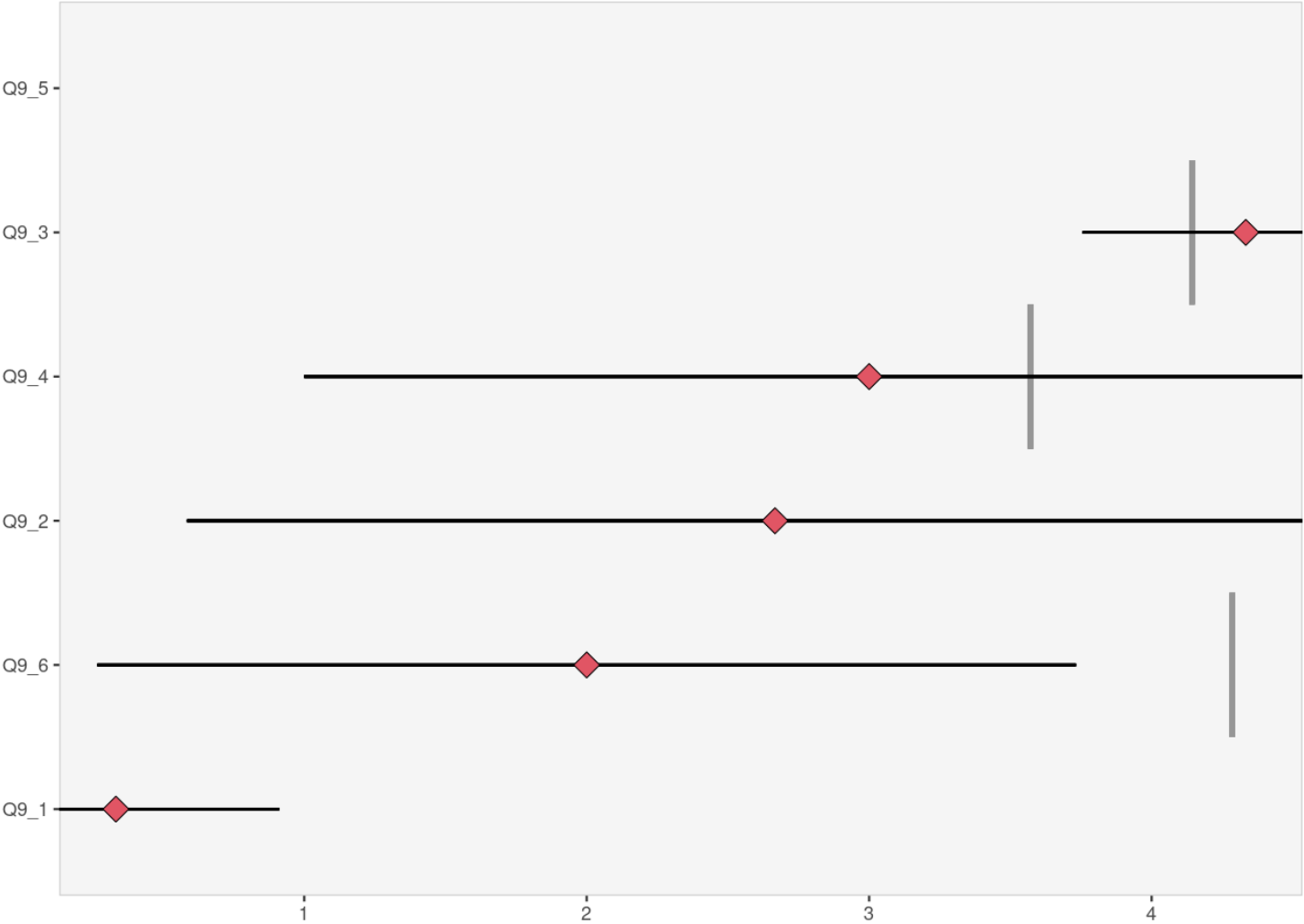


Spider chart.

Segment 1 profile

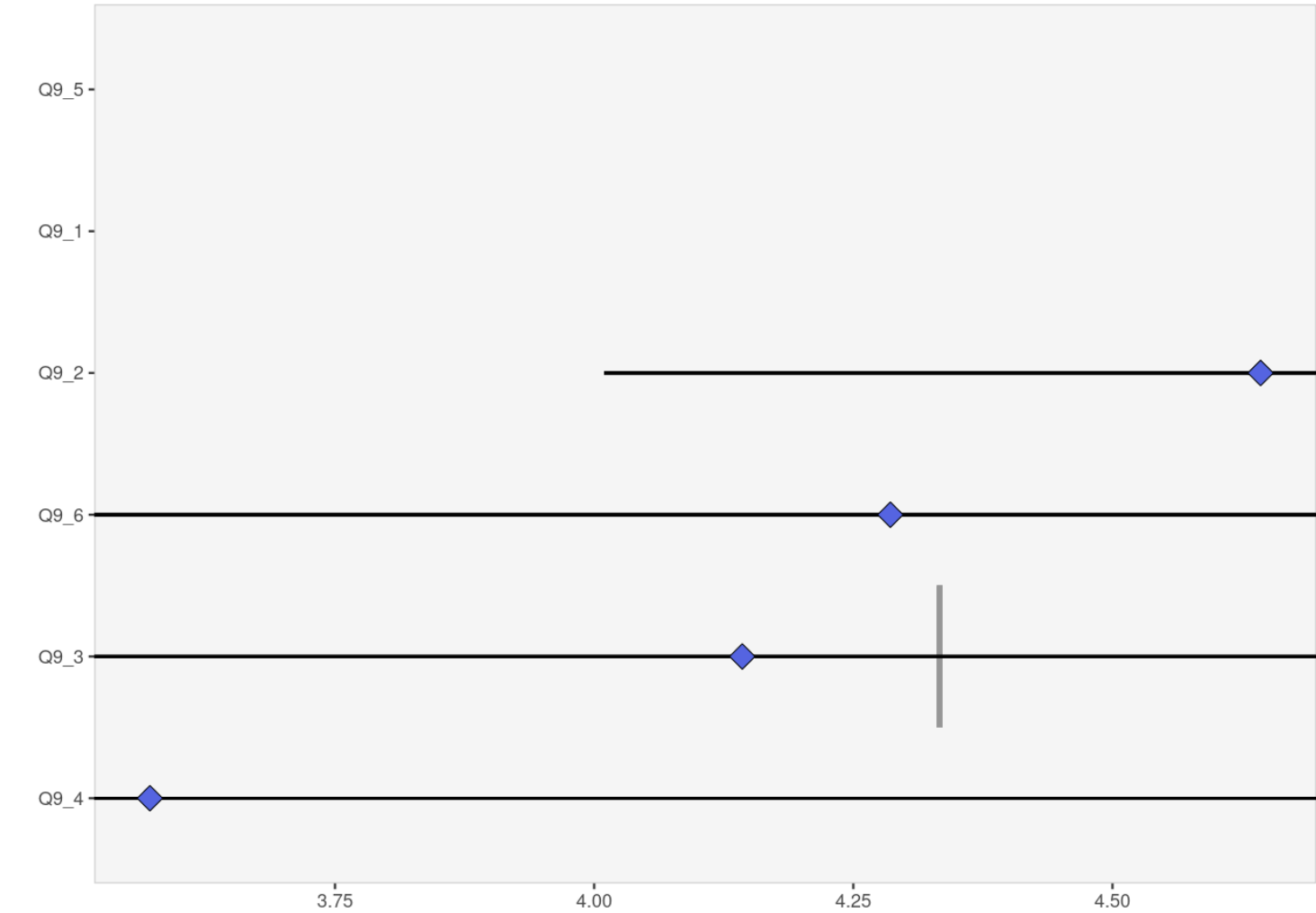
The following charts represent the profile of each segment. These charts are only available when the data are not standardized, hence the model assumes that all segmentation variables use the same scale.

- For each segment, the segmentation variables are ordered in decreasing order of magnitude.
- The colored dots represent the average of the segment.
- The horizontal lines represent the standard deviations within that segment.
- The vertical, gray lines represent the averages of the rest of the population, after excluding members of the segment under scrutiny.



Segment 1 profile.

Segment 2 profile



Segment 2 profile.

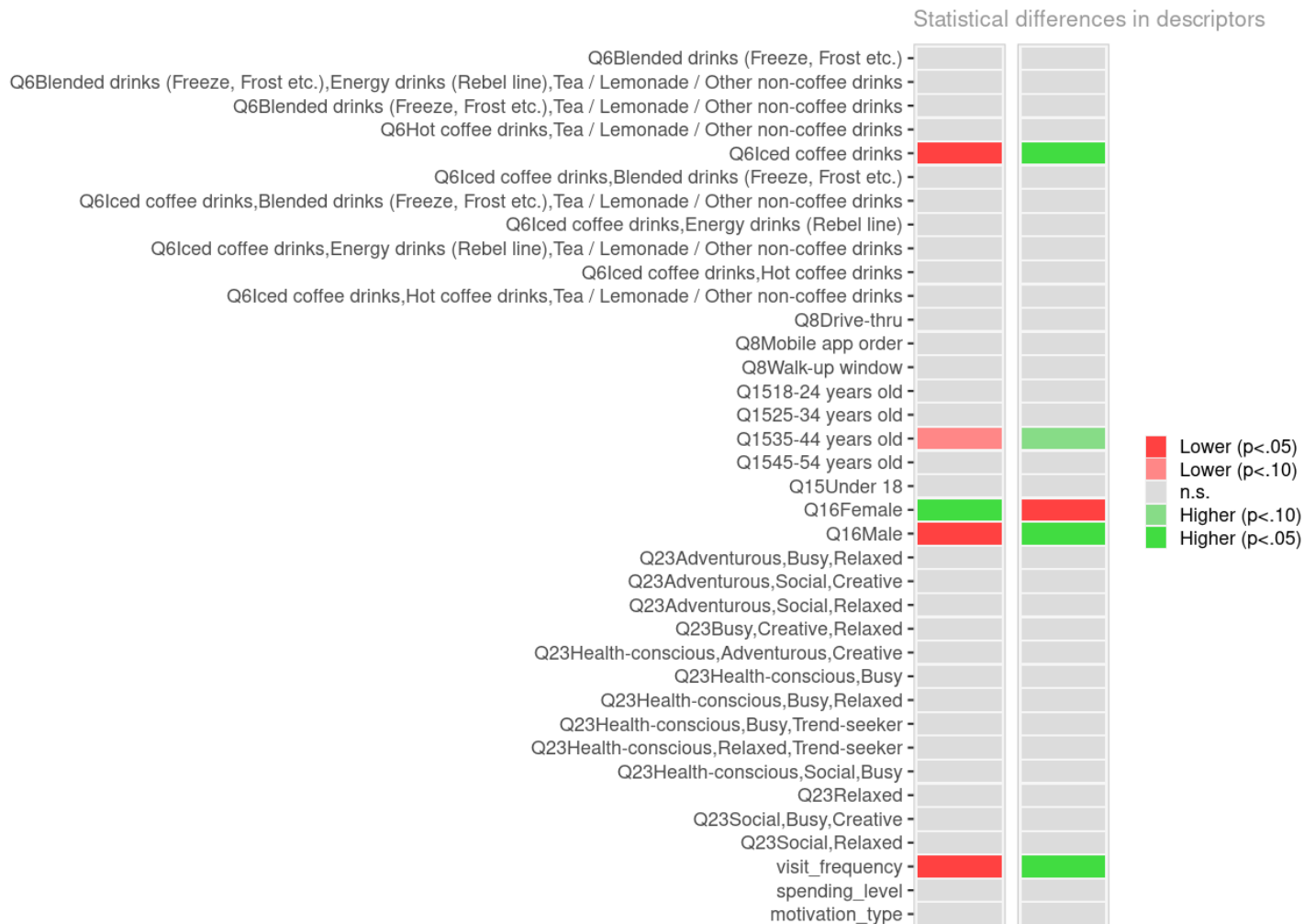
Descriptor analysis

Descriptors

This table reports the descriptor averages of each segment. The more differences can be found, the easier it will be to predict segment membership based on descriptors alone.

	Population	Segment 1	Segment 2
Q6Blended drinks (Freeze, Frost etc.)	0.118	0	0.143
Q6Blended drinks (Freeze, Frost etc.),Energy drinks (Rebel line),Tea / Lemonade / Other non-coffee drinks	0.059	0.333	0
Q6Blended drinks (Freeze, Frost etc.),Tea / Lemonade / Other non-coffee drinks	0.059	0	0.071
Q6Hot coffee drinks,Tea / Lemonade / Other non-coffee drinks	0.059	0	0.071
Q6Iced coffee drinks	0.294	0	0.357
Q6Iced coffee drinks,Blended drinks (Freeze, Frost etc.)	0.059	0	0.071
Q6Iced coffee drinks,Blended drinks (Freeze, Frost etc.),Tea / Lemonade / Other non-coffee drinks	0.059	0	0.071
Q6Iced coffee drinks,Energy drinks (Rebel line)	0.059	0	0.071
Q6Iced coffee drinks,Energy drinks (Rebel line),Tea / Lemonade / Other non-coffee drinks	0.059	0	0.071
Q6Iced coffee drinks,Hot coffee drinks	0.059	0.333	0
Q6Iced coffee drinks,Hot coffee drinks,Tea / Lemonade / Other non-coffee drinks	0.118	0.333	0.071
Q8Drive-thru	0.647	0.667	0.643
Q8Mobile app order	0.118	0	0.143
Q8Walk-up window	0.235	0.333	0.214
Q1518-24 years old	0.353	0.667	0.286
Q1525-34 years old	0.353	0.333	0.357
Q1535-44 years old	0.176	0	0.214
Q1545-54 years old	0.059	0	0.071
Q15Under 18	0.059	0	0.071
Q16Female	0.765	1	0.714
Q16Male	0.235	0	0.286
Q23Adventurous,Busy,Relaxed	0.059	0	0.071
Q23Adventurous,Social,Creative	0.059	0.333	0
Q23Adventurous,Social,Relaxed	0.059	0	0.071
Q23Busy,Creative,Relaxed	0.059	0	0.071
Q23Health-conscious,Adventurous,Creative	0.059	0	0.071
Q23Health-conscious,Busy	0.059	0	0.071
Q23Health-conscious,Busy,Relaxed	0.059	0	0.071
Q23Health-conscious,Busy,Trend-seeker	0.059	0	0.071
Q23Health-conscious,Relaxed,Trend-seeker	0.059	0	0.071
Q23Health-conscious,Social,Busy	0.176	0.667	0.071
Q23Relaxed	0.059	0	0.071
Q23Social,Busy,Creative	0.118	0	0.143
Q23Social,Relaxed	0.118	0	0.143
visit_frequency	1.41	1	1.5
spending_level	NA	1.67	NA
motivation_type	1.71	2	1.64

Descriptor data per segment. Average value of each descriptor, overall and within each cluster. Descriptors that are statistically different from the rest of the population are highlighted in red (lower) or green (higher).



Descriptor differences per segment. Cell colors indicate to what extent the distribution of a descriptor in a segment is statistically different from the rest of the population.

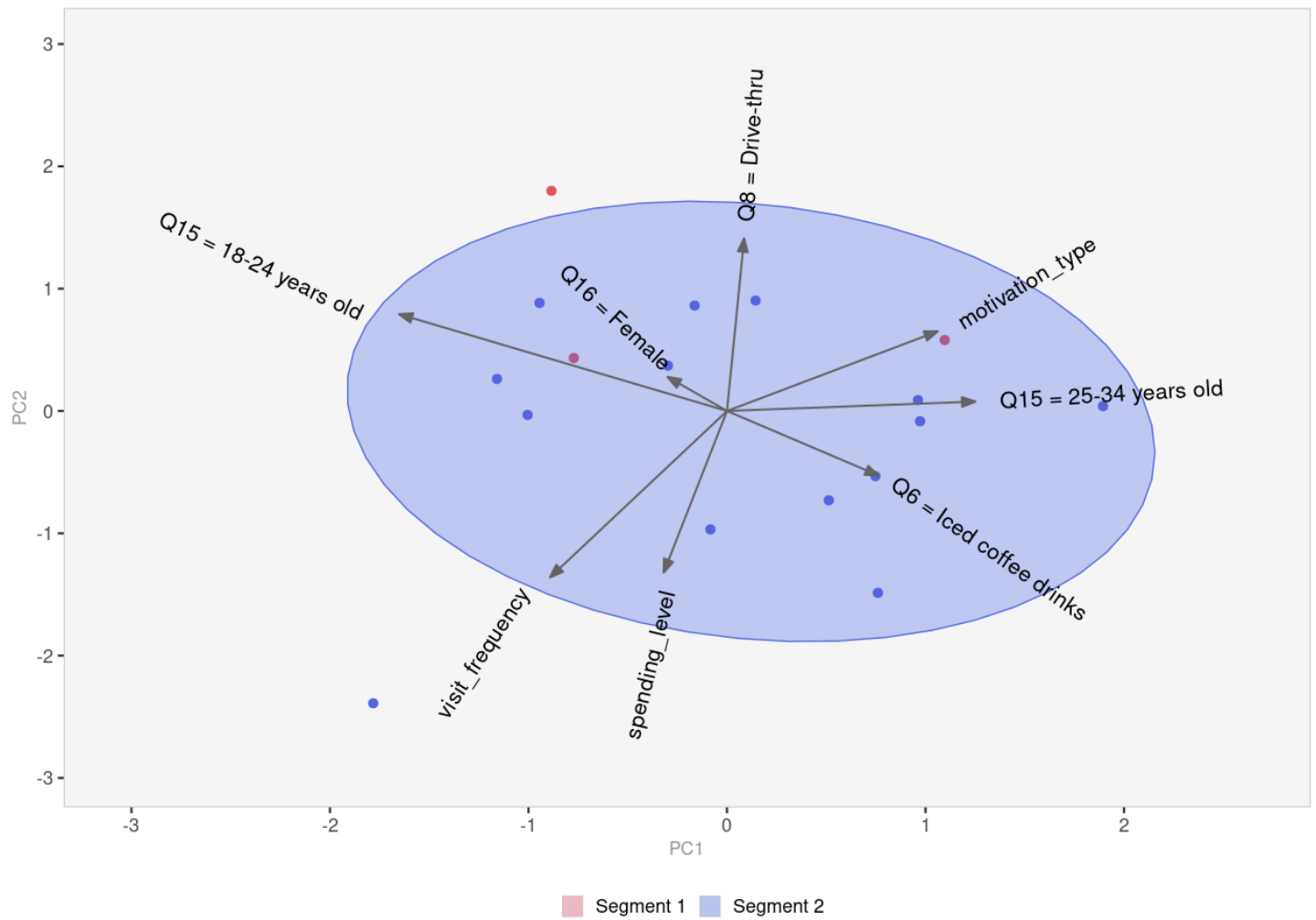
Descriptor space

The chart below is a graphical representation of the various segments, segment members, and descriptors. It is obtained by outputting the first two dimensions of a principal component analysis performed on the (standardized) descriptors, on top of which segment information has been overlaid.

Because only the first two dimensions of the PCA are displayed, and these two dimensions capture only 48.9% of the variance in the data, some differences between segments might not appear here. Note that descriptors with no variance, if any, have been excluded.

If two or more segments fully overlap, it is unlikely that they could be clearly separated based on descriptors alone.

However, two segments that seem to overlap on two dimensions may be more clearly separated on other dimensions. Consequently, the confusion matrix is a better guide to assess the quality of segment classification based on descriptors.



Descriptor space. Spatial representation of segments and their descriptors, using principal component analysis.

Classification model

Introduction

Often, segmentation (needs) variables for each customer may not be available to managers, but descriptors variables for customers may be available.

In this section, we explore whether descriptors alone can predict segment membership with sufficient accuracy. The confusion matrix and hit rates (reported below) indicate whether the model is accurate enough.

For member classification based on descriptors, Enginius uses a multinomial logit model (similar to the one used to predict 'choices between multiple alternatives (A/B/C)' in the predictive modeling module.

The largest segment is selected as the default option (dummy), and the model identifies which descriptors are the most significant for predicting cluster memberships. If a descriptor is highly predictive, its p-values will be close to zero, and the cells will appear in green (or red).

Model coefficients

	Segment summary(model)\$coefficients
(Intercept)	14
Q6 = Iced coffee drinks	-67
Q8 = Drive-thru	18.4
Q15 = 18-24 years old	39.4
Q15 = 25-34 years old	-15.7
Q16 = Female	26.3
visit_frequency	-30.4
spending_level	-57.1
motivation_type	36.9

Model parameters. Segment 2 is the model baseline.

P-values

	Segment summary(model)\$coefficients
(Intercept)	0.838
Q6 = Iced coffee drinks	NaN
Q8 = Drive-thru	0.896
Q15 = 18-24 years old	0.748
Q15 = 25-34 years old	0.933
Q16 = Female	0.703
visit_frequency	0.659
spending_level	0.678
motivation_type	0.856

p-values. Probabilities that parameter estimates are different from zero only by chance.

Confusion matrix

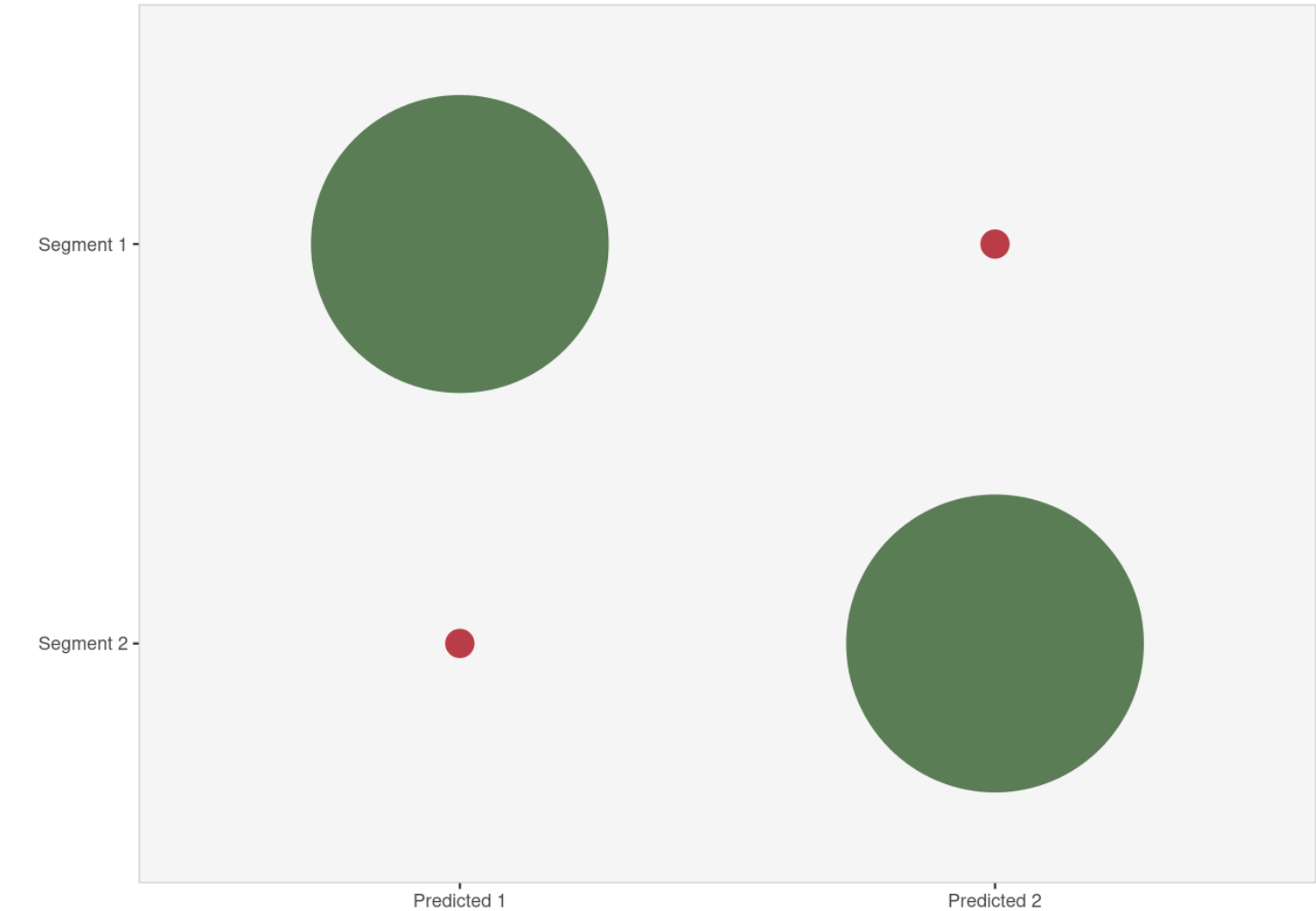
The confusion matrix compares actual segment membership (obtained from the segmentation analysis and the original segmentation variables) and predicted segment membership (obtained from the in-sample classification analysis and the descriptors alone). When actual and predicted segment memberships coincide, the diagonal elements will be comparatively large, indicating that the classification model based on available descriptors is accurate.

	Predicted 1	Predicted 2	Total
Segment 1	3	0	3
Segment 2	0	14	14
Total	3	14	17

Confusion matrix (count). The model has correctly classified 17 of the 17 observations. The off-diagonal elements are classification errors.

	Predicted 1	Predicted 2	Total
Segment 1	100%	0%	100%
Segment 2	0%	100%	100%

Confusion matrix (%). The global hit rate of the model is 100%. The diagonal elements represent segment-specific hit rates.



Confusion matrix (plot). Graphic representation of the confusion matrix: actual segment membership versus predicted segment membership. Bubbles in the diagonale represent correct classification.

Model predictions

	Prob(cluster 1)	Prob(cluster 2)	Predicted	Actual	Correct
1	100%	0%	1	1	1
2	0%	100%	2	2	1
3	0%	100%	2	2	1
4	0%	100%	2	2	1
5	0%	100%	2	2	1
6	0%	100%	2	2	1
7	0%	100%	2	2	1
8	0%	100%	2	2	1
9	100%	0%	1	1	1
10	0%	100%	2	2	1

Model predictions (in-sample) (excerpt). This table details the probabilities of each member of the segmentation dataset to belong to each cluster (as predicted by the in-sample classification model and the descriptors alone). The segment with the highest probability is retained, and is compared to the actual segment membership to measure model accuracy and classification errors.

Copyright (c) 2025, DecisionPro Inc.