enginius
MARKETING ENGINEERING ONLINE

Enginius

Segmentation

Sri Harsha Kommineni, University of Tampa

Table of Contents

Warnings

Segmentation options

Data transformation

Segment solution

Segment description

Segment profiles

Descriptor analysis

Classification model

Out-of-sample classification

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 - frequency of Purchase = Once a month or more, Gender = Non-binary/Other, Primary purpose = Gym, and back strain prevention, spend on purchasing = More than $250

Segmentation options

Options selected

Option	Selection
Clustering method	Hierarchical
Standardization method	byrow
Segments forced	No
Run discriminant analysis	Yes
Run classification analysis	Yes
Date and time	2024-11-22 00:43:10 UTC

Options selected.

Data description

	Data	Number of Rows	Number of columns	Column names
1	Segmentation data	131	7	\, Innovation, Comfort, Style / Fashion Appeal, Brand Trustworthiness, ...
2	Discriminant data	131	7	\, Age, Gender, Annual income level, frequency of Purchase, ...
3	Clssification data	131	7	\, Age, Gender, Annual income level, frequency of Purchase, ...

Data description.

Data transformation

The segmentation data has been scaled row wise

	Mean	Standard deviation
Respondent 1	4.000	0.6325
Respondent 2	4.000	0.8944
Respondent 3	4.167	0.7528
Respondent 4	2.167	0.9832
Respondent 5	4.000	1.2649
Respondent 6	3.000	0.0000
Respondent 7	4.333	0.8165
Respondent 8	2.000	0.6325
Respondent 9	4.167	0.7528
Respondent 10	4.167	1.1690

Mean and standard deviation row wise (excerpt).

Segment solution

3-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.

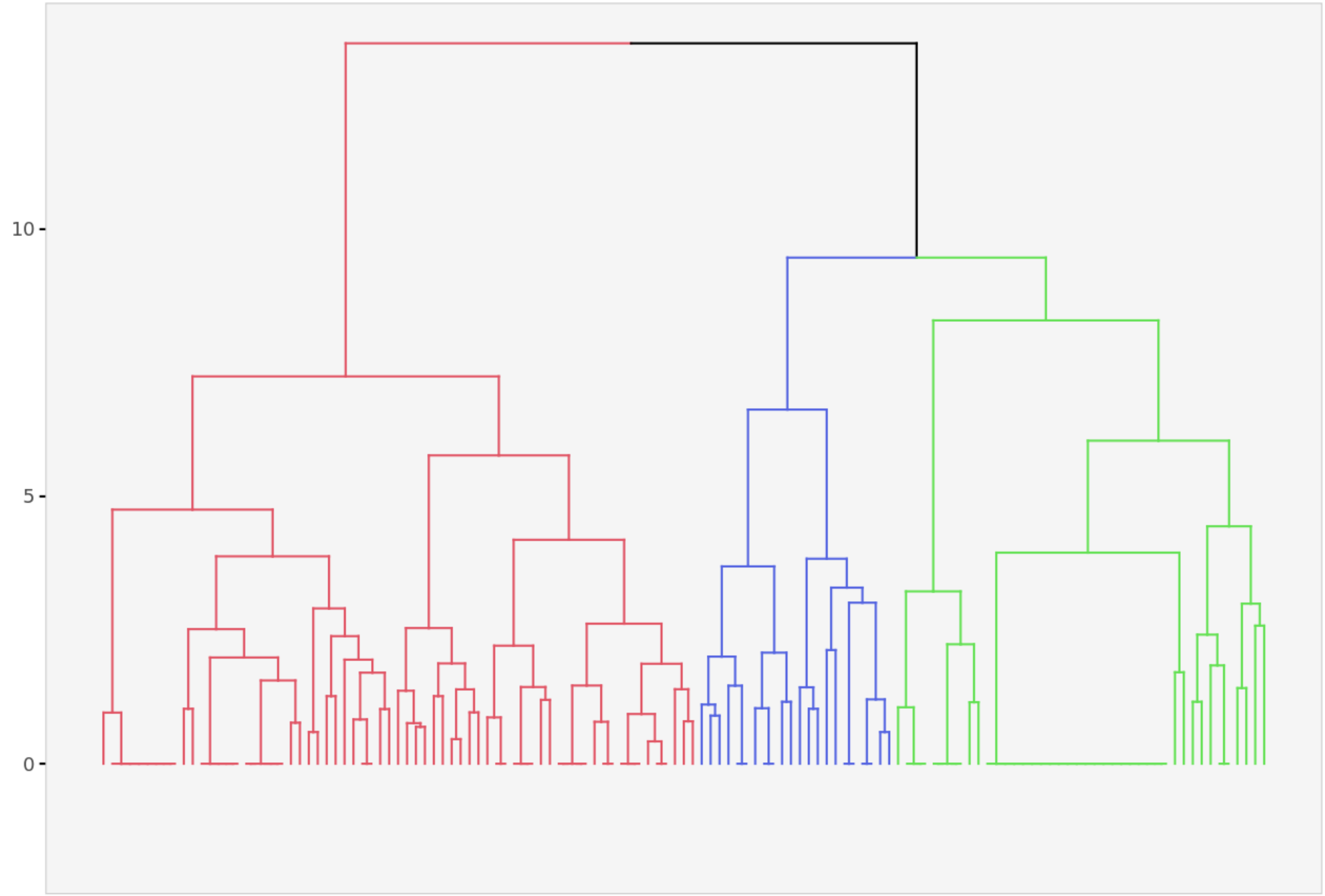
Using a statistical criteria exclusively (see scree plot analysis below), we have retained 3 segments.

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.



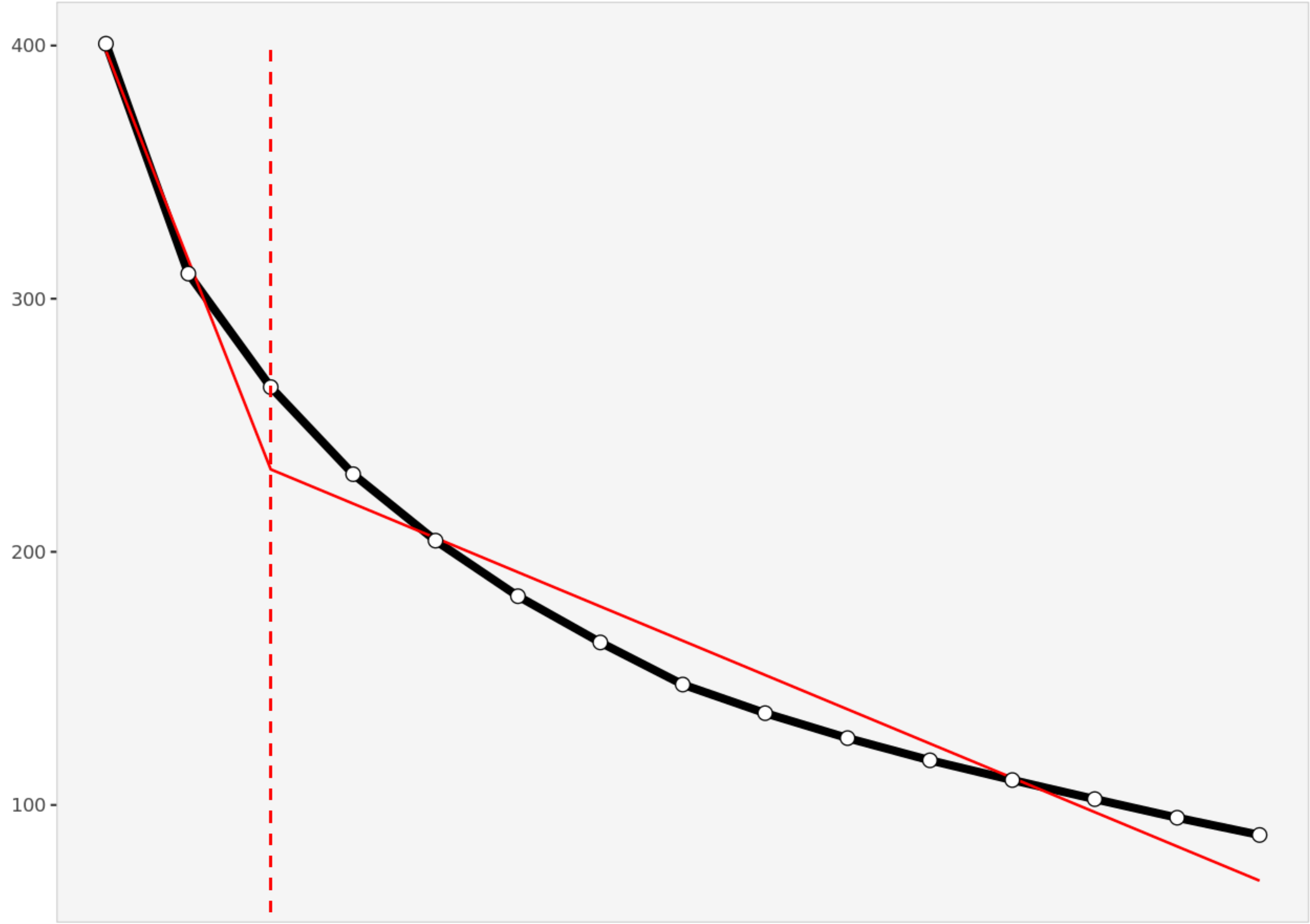
Dendrogram. The dendrogram is a tree diagram to illustrate the arrangement of clusters produced by hierarchical clustering, and how the observations are incrementally clustered together.

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 maxumum 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

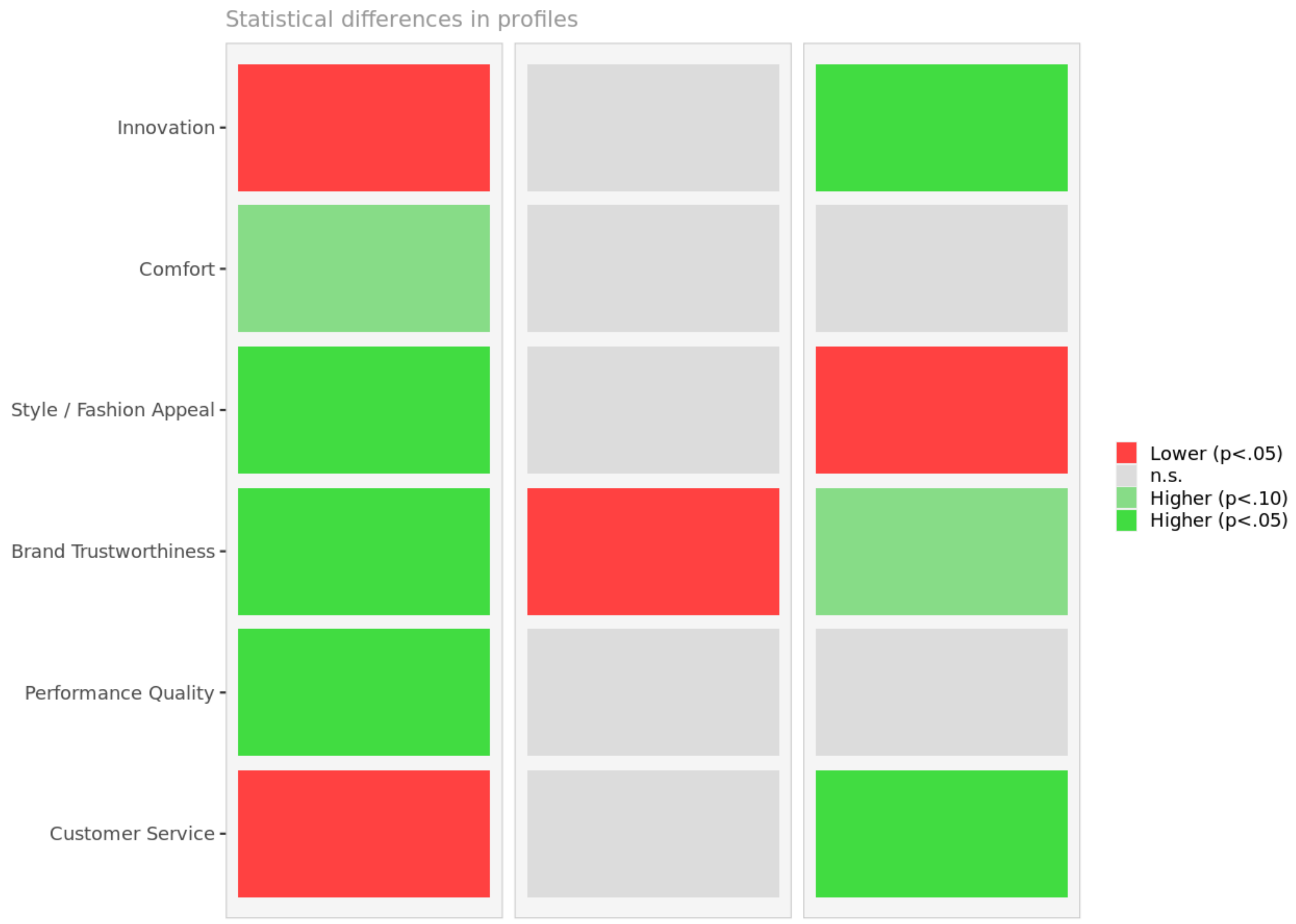
	Population	Segment 1	Segment 2	Segment 3
Size	131	67	22	42
Relative size	100%	51%	17%	32%

Segment size.

Segment description

	Population	Segment 1	Segment 2	Segment 3
Innovation	2.77	2.28	2.82	3.52
Comfort	3.91	4.06	3.82	3.71
Style / Fashion Appeal	3.68	3.96	3.41	3.38
Brand Trustworthiness	3.47	3.72	2.18	3.74
Performance Quality	3.72	3.96	3.41	3.50
Customer Service	3.02	2.69	2.86	3.64

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).



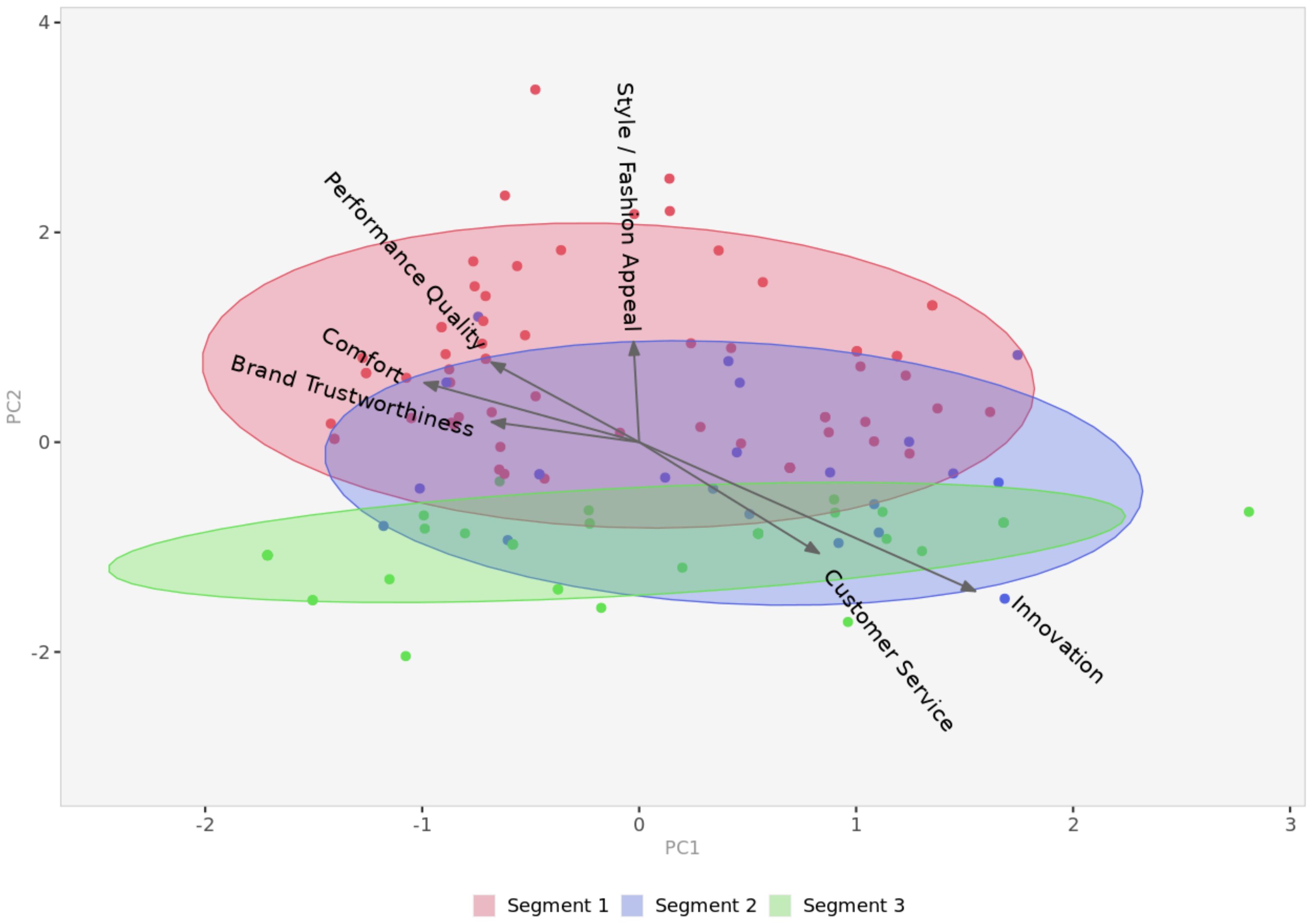
Segment differences per segment. Cell colors indicate to what extent a segment is statistically different from the rest of the population on each segmentation variable.

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 79.8% 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

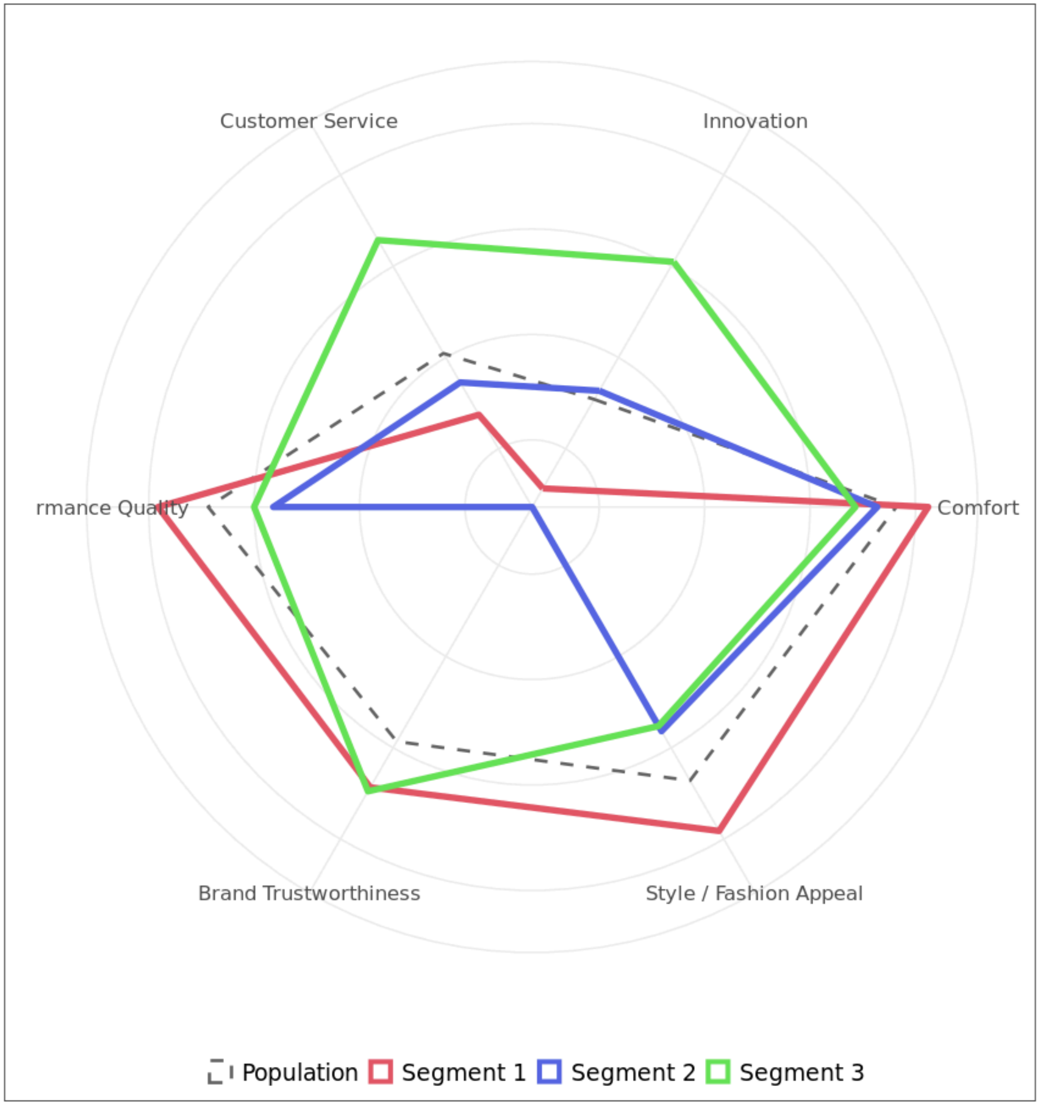
	Segment
Respondent 1	1
Respondent 2	3
Respondent 3	1
Respondent 4	2
Respondent 5	2
Respondent 6	3
Respondent 7	1
Respondent 8	1
Respondent 9	1
Respondent 10	1

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.

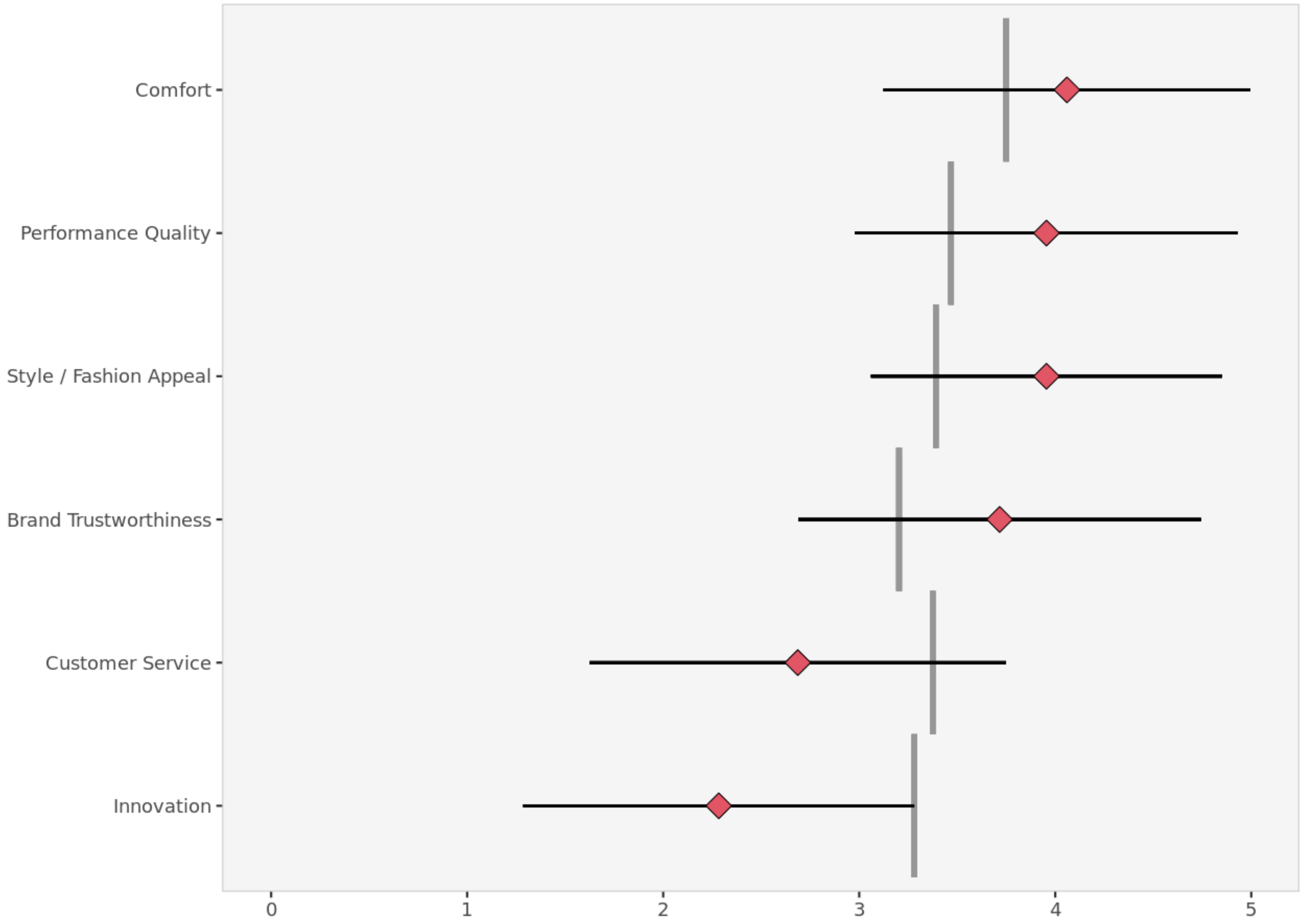


Spider chart.

Segment 1 profile

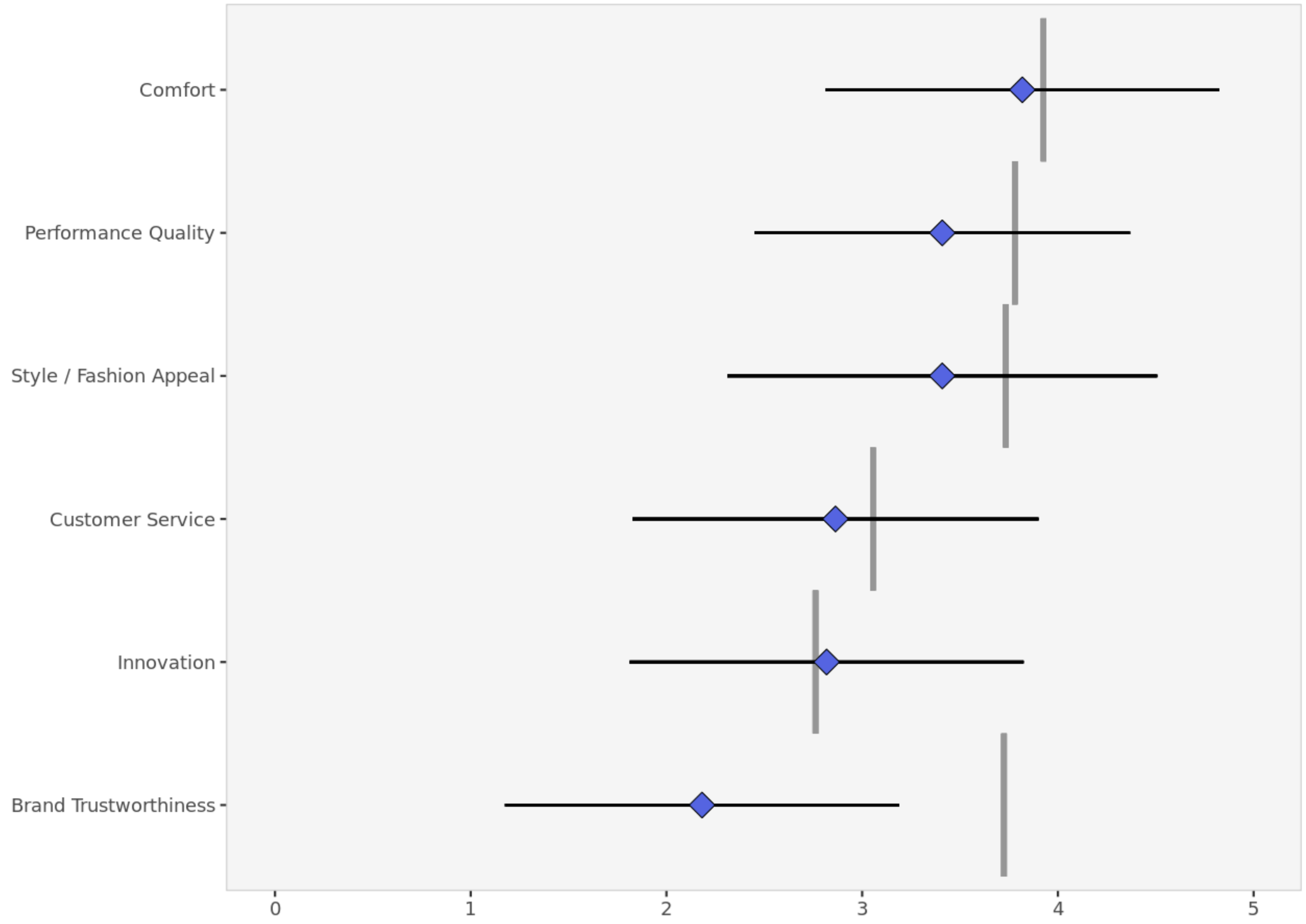
The following charts represent the profile of each segment. These charts are only available when the data are not standardized, hence the model assumes that all segmentation variables use the same scale.

- 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.



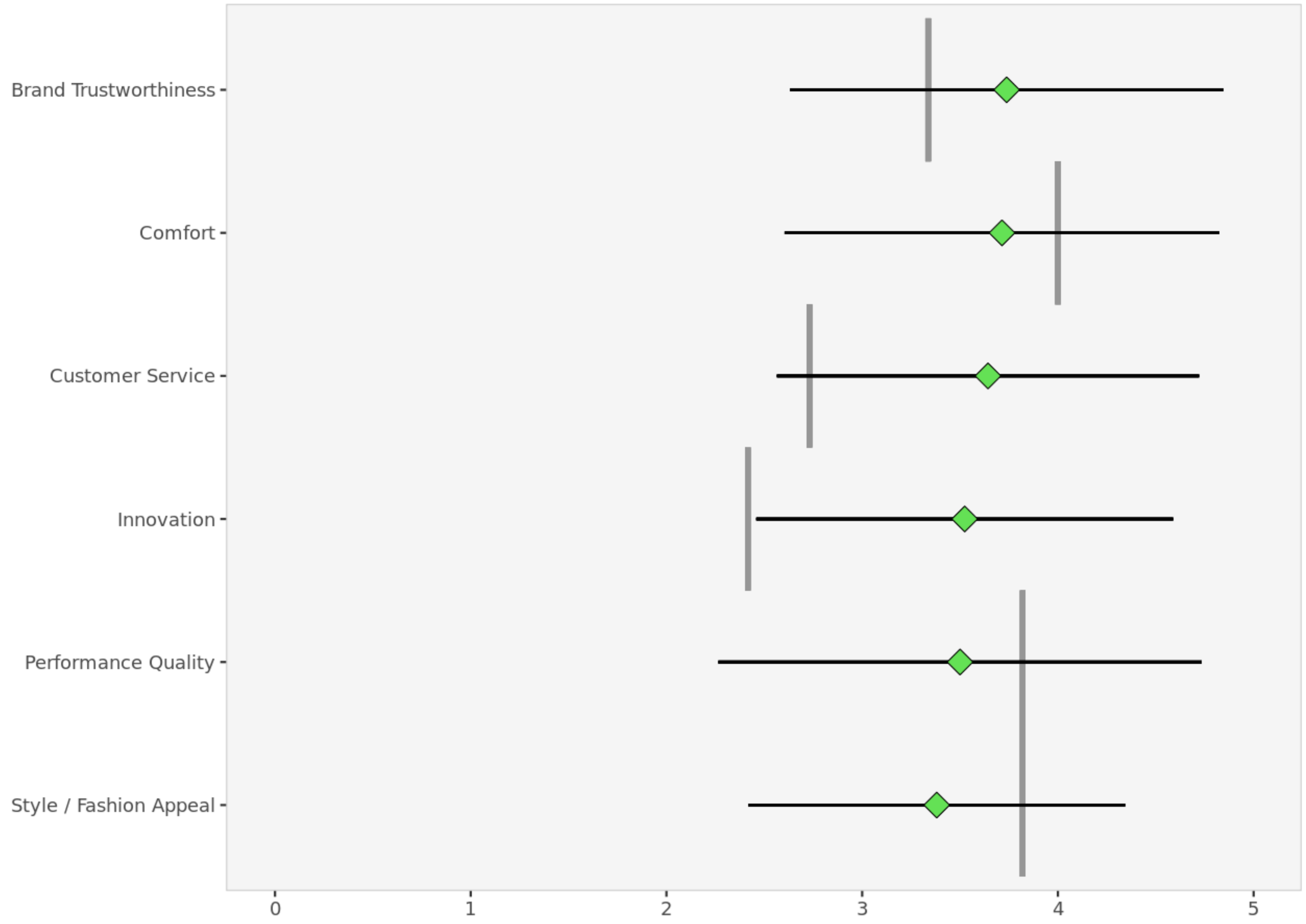
Segment 1 profile.

Segment 2 profile



Segment 2 profile.

Segment 3 profile



Segment 3 profile.

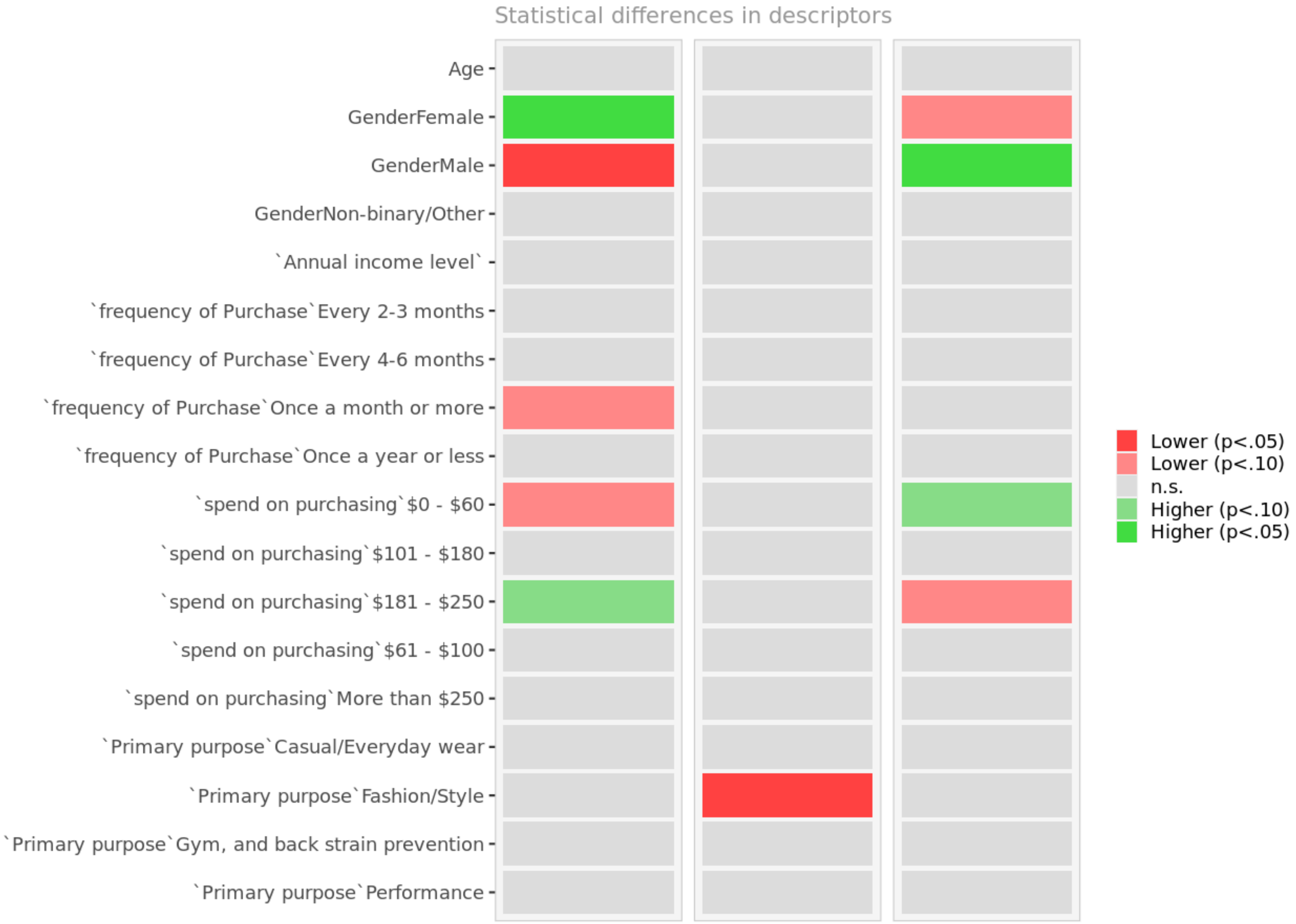
Descriptor analysis

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	Segment 3
Age	23.7	23.4	24.7	23.8
GenderFemale	0.664	0.746	0.636	0.548
GenderMale	0.328	0.239	0.364	0.452
GenderNon-binary/Other	0.008	0.015	0.000	0.000
`Annual income level`	22.1	20.2	29.5	21.2
`frequency of Purchase`Every 2-3 months	0.198	0.239	0.136	0.167
`frequency of Purchase`Every 4-6 months	0.282	0.313	0.318	0.214
`frequency of Purchase`Once a month or more	0.160	0.104	0.227	0.214
`frequency of Purchase`Once a year or less	0.359	0.343	0.318	0.405
`spend on purchasing`$0 - $60	0.107	0.060	0.091	0.190
`spend on purchasing`$101 - $180	0.412	0.388	0.409	0.452
`spend on purchasing`$181 - $250	0.115	0.164	0.091	0.048
`spend on purchasing`$61 - $100	0.305	0.313	0.318	0.286
`spend on purchasing`More than $250	0.061	0.075	0.091	0.024
`Primary purpose`Casual/Everyday wear	0.527	0.582	0.500	0.452
`Primary purpose`Fashion/Style	0.107	0.119	0.000	0.143
`Primary purpose`Gym, and back strain prevention	0.008	0.000	0.045	0.000
`Primary purpose`Performance	0.359	0.299	0.455	0.405

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.

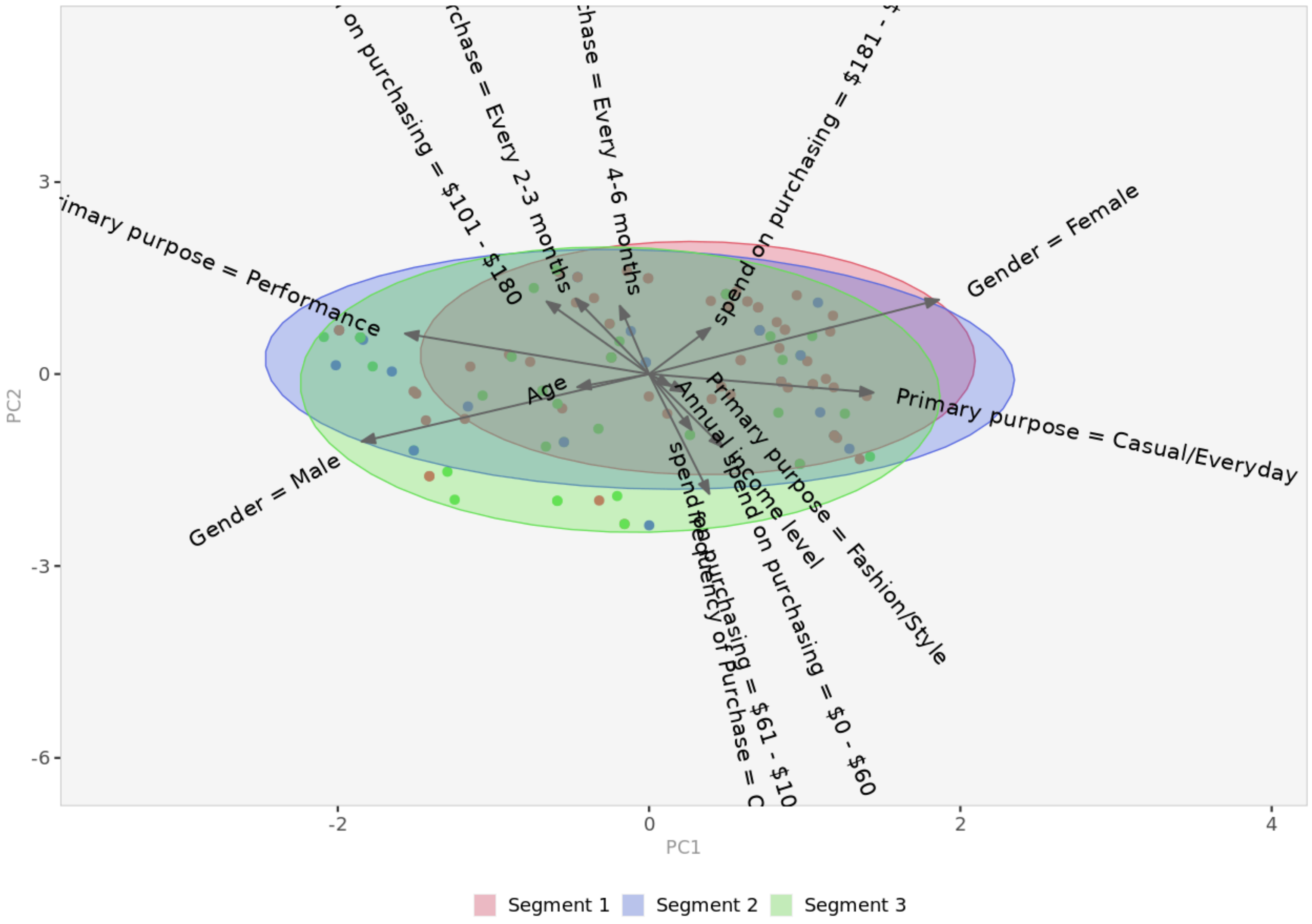
Descriptor space

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

Because only the first two dimensions of the PCA are displayed, and these two dimensions capture only 30.9% 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 2	Segment 3
(Intercept)	8.34	-21.40
Age	0.043	0.039
Annual income level	0.015	0.006
frequency of Purchase = Once a year or less	-1.219	-0.939
frequency of Purchase = Every 4-6 months	-1.10	-1.21
frequency of Purchase = Every 2-3 months	-2.52	-1.19
Gender = Male	16.3	14.1
Gender = Female	16.3	13.3
Primary purpose = Performance	-24.98	6.33
Primary purpose = Casual/Everyday wear	-25.89	5.61
Primary purpose = Fashion/Style	-41.76	6.04
spend on purchasing = $61 - $100	-0.565	1.281
spend on purchasing = $101 - $180	-0.185	1.456
spend on purchasing = $181 - $250	-1.287	-0.036
spend on purchasing = $0 - $60	0.194	2.470

Model parameters. Segment 1 is the model baseline.

P-values

	Segment 2	Segment 3
(Intercept)	0.000	0.000
Age	0.390	0.403
Annual income level	0.100	0.507
frequency of Purchase = Once a year or less	0.121	0.154
frequency of Purchase = Every 4-6 months	0.175	0.090
frequency of Purchase = Every 2-3 months	0.021	0.122
Gender = Male	0.000	0.000
Gender = Female	0.000	0.000
Primary purpose = Performance	0.000	0.000
Primary purpose = Casual/Everyday wear	0.000	0.000
Primary purpose = Fashion/Style	0.000	0.000
spend on purchasing = $61 - $100	0.601	0.305
spend on purchasing = $101 - $180	0.856	0.232
spend on purchasing = $181 - $250	0.320	0.980
spend on purchasing = $0 - $60	0.886	0.073

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

Confusion matrix

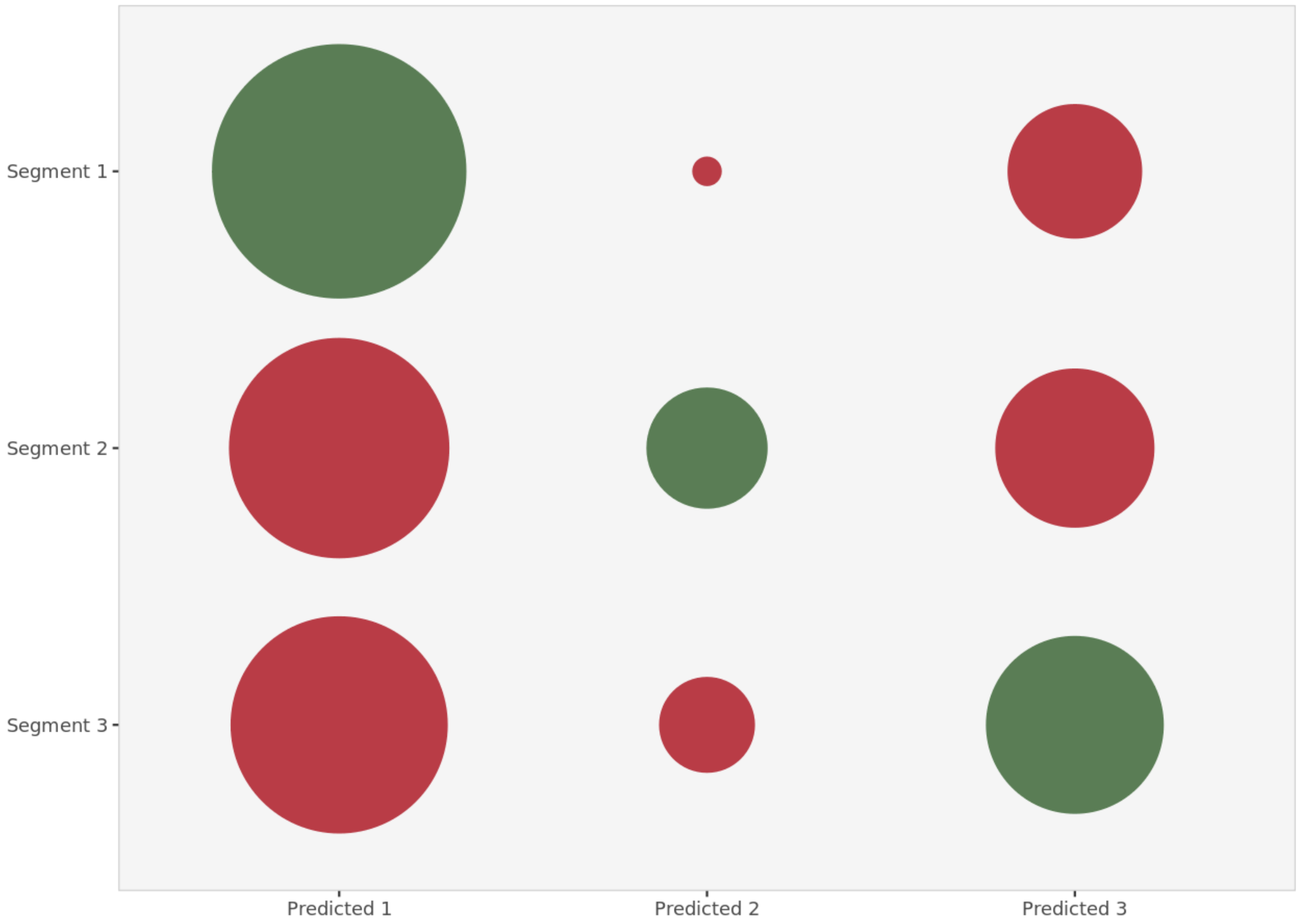
The confusion matrix compares actual segment membership (obtained from the segmentation analysis and the original segmentation variables) and predicted segment membership (obtained from the in-sample classification analysis and the descriptors alone). When actual and predicted segment memberships coincide,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

	Predicted 1	Predicted 2	Predicted 3	Total
Segment 1	55	0	12	67
Segment 2	13	3	6	22
Segment 3	24	3	15	42
Total	92	6	33	131

Confusion matrix (count). The model has correctly classified 73 of the 131 observations. The off-diagonal elements are classification errors.

	Predicted 1	Predicted 2	Predicted 3	Total
Segment 1	82%	0%	18%	100%
Segment 2	59%	14%	27%	100%
Segment 3	57%	7%	36%	100%

Confusion matrix (%). The global hit rate of the model is 56%. 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)	Prob(cluster 3)	Predicted	Actual	Correct
Respondent 1	31%	16%	53%	3	1	0
Respondent 2	65%	16%	20%	1	3	0
Respondent 3	61%	4%	36%	1	1	1
Respondent 4	52%	23%	25%	1	2	0
Respondent 5	63%	14%	23%	1	2	0
Respondent 6	50%	0%	50%	1	3	0
Respondent 7	70%	18%	13%	1	1	1
Respondent 8	54%	0%	46%	1	1	1
Respondent 9	65%	17%	18%	1	1	1
Respondent 10	65%	12%	23%	1	1	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.

Out-of-sample classification

Introduction

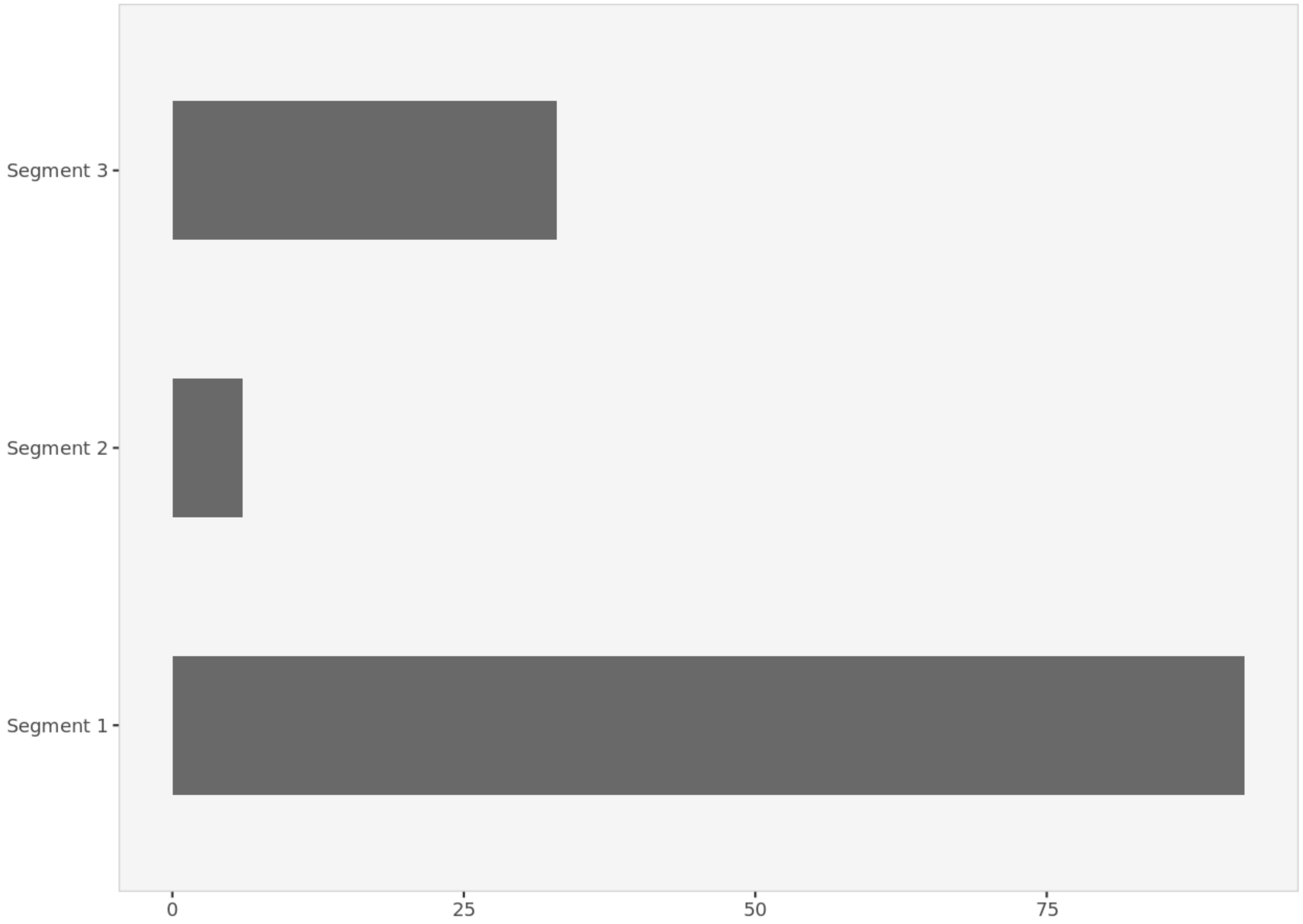
In this section, we apply the classification model (calibrated above) to a new population of never-befeore-seen contacts, for whom only descriptors are available.

Because segmentation variables and actual segment membership are unavailable, the actual accuracy of the model predictions are unknown and can only be inferred from the previous section.

Segment size

	Population	Segment 1	Segment 2	Segment 3
Size	131	92	6	33
Relative size	100%	70%	5%	25%

Segment size (table).



Segment size (chart).

Model predictions

	Prob(cluster 1)	Prob(cluster 2)	Prob(cluster 3)	Predicted
Respondent 1	31%	16%	53%	3
Respondent 2	65%	16%	20%	1
Respondent 3	61%	4%	36%	1

Respondent 4	52%	23%	25%	1
Respondent 5	63%	14%	23%	1
Respondent 6	50%	0%	50%	1
Respondent 7	70%	18%	13%	1
Respondent 8	54%	0%	46%	1
Respondent 9	65%	17%	18%	1
Respondent 10	65%	12%	23%	1

Model predictions (out-sample) (excerpt). This table details the probabilities of each member of the out-of-sample classification data to belong to each cluster (as predicted by the classification model and the descriptors alone). The segment with the highest probability is retained.