enginius
MARKETING ENGINEERING ONLINE

Enginius

Segmentation

Yu Chin Chen, Arizona State University

Table of Contents

Segmentation options

Options selected

Option	Selection
Clustering method	K-means
Standardization method	none
Segments forced	3
Run discriminant analysis	Yes
Run classification analysis	Yes
Date and time	2025-02-13 21:20:33 UTC

Options selected.

Data description

	Data	Number of Rows	Number of columns	Column names
1	Segmentation data	317	22	Id, Rich full-bodied, Light beer, No aftertaste, Refreshing, ...
2	Discriminant data	317	13	Id, Weekly consumption, Age (1-7), Income (1-7), Education (1-6), ...

Data description.

Data transformation

Standardization has not been performed.

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.

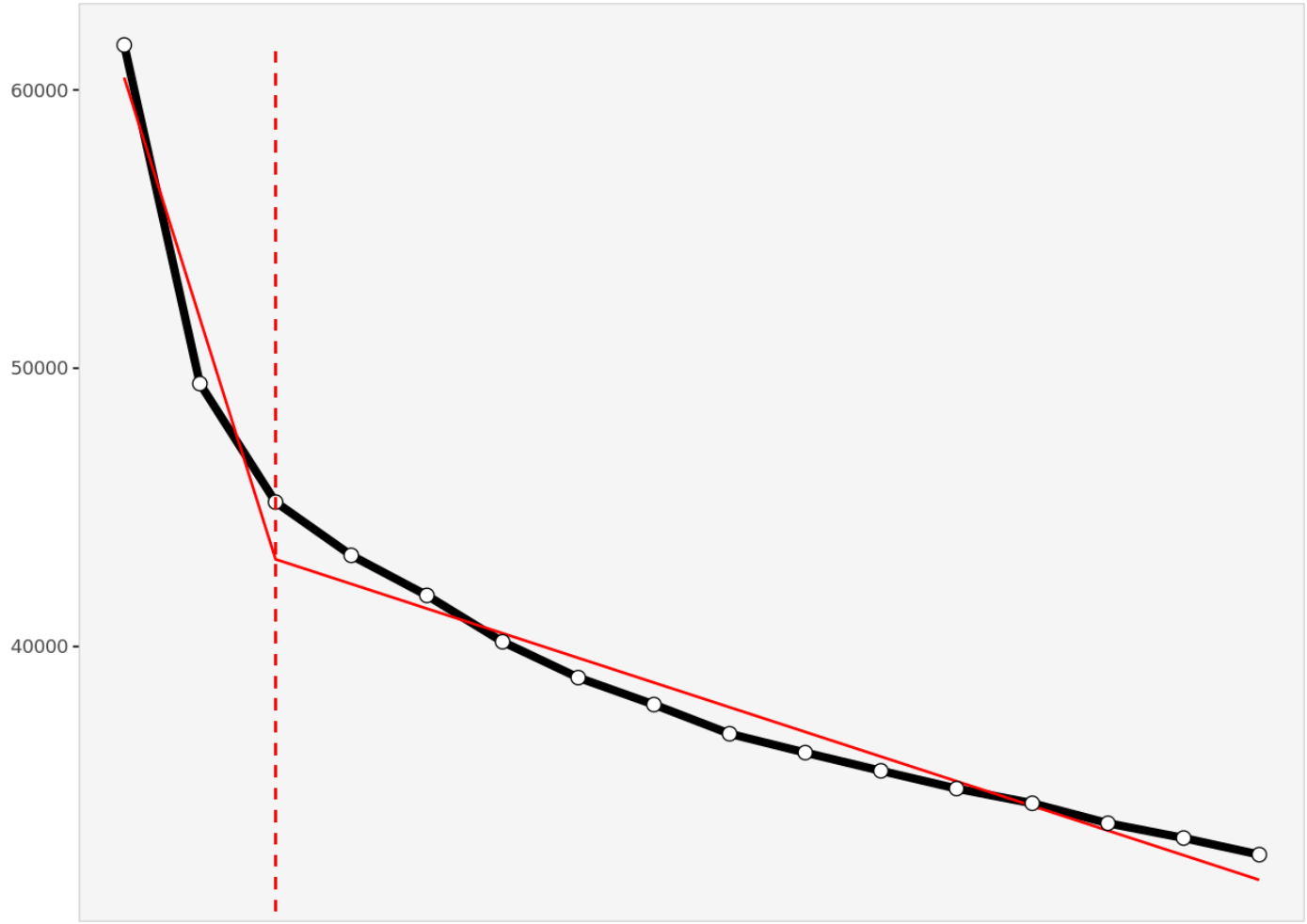
You have decided to perform the analysis with 3 segments.

The segmentation method relies on the k-means approach. This approach does not generate a dendrogram.

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.



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	317	119	108	90
Relative size	100%	38%	34%	28%

Segment size.

Segment description

	Population	Segment 1	Segment 2	Segment 3
Rich full-bodied	4.77	4.40	7.29	2.22
Light beer	3.72	3.92	3.98	3.14
No aftertaste	4.56	5.28	5.36	2.66
Refreshing	5.02	5.09	7.19	2.33
Goes down easily	5.17	5.72	6.70	2.59
Gives a buzz	3.39	3.27	3.89	2.97
Good taste	3.000	0.176	8.306	0.367
Low price	3.91	4.16	4.66	2.69
Good value	4.65	4.48	5.83	3.46
From country with brewing tradition	3.82	3.86	4.41	3.06
Attractive bottle	3.00	2.81	3.33	2.84
Prestigious brand	3.20	3.12	3.94	2.41
High quality	4.48	4.58	7.10	1.19
Drink at picnics	4.56	5.29	5.34	2.67
Masculine	2.67	2.35	3.44	2.17
For young people	2.49	2.11	3.16	2.19
Drink with friends	4.70	5.93	5.91	1.61
Drink at home	4.34	5.25	5.63	1.58
To serve dinner guests	4.92	5.91	6.44	1.79
For dining out	5.03	5.97	6.27	2.31
Drink at bar	4.35	4.77	5.55	2.34

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).

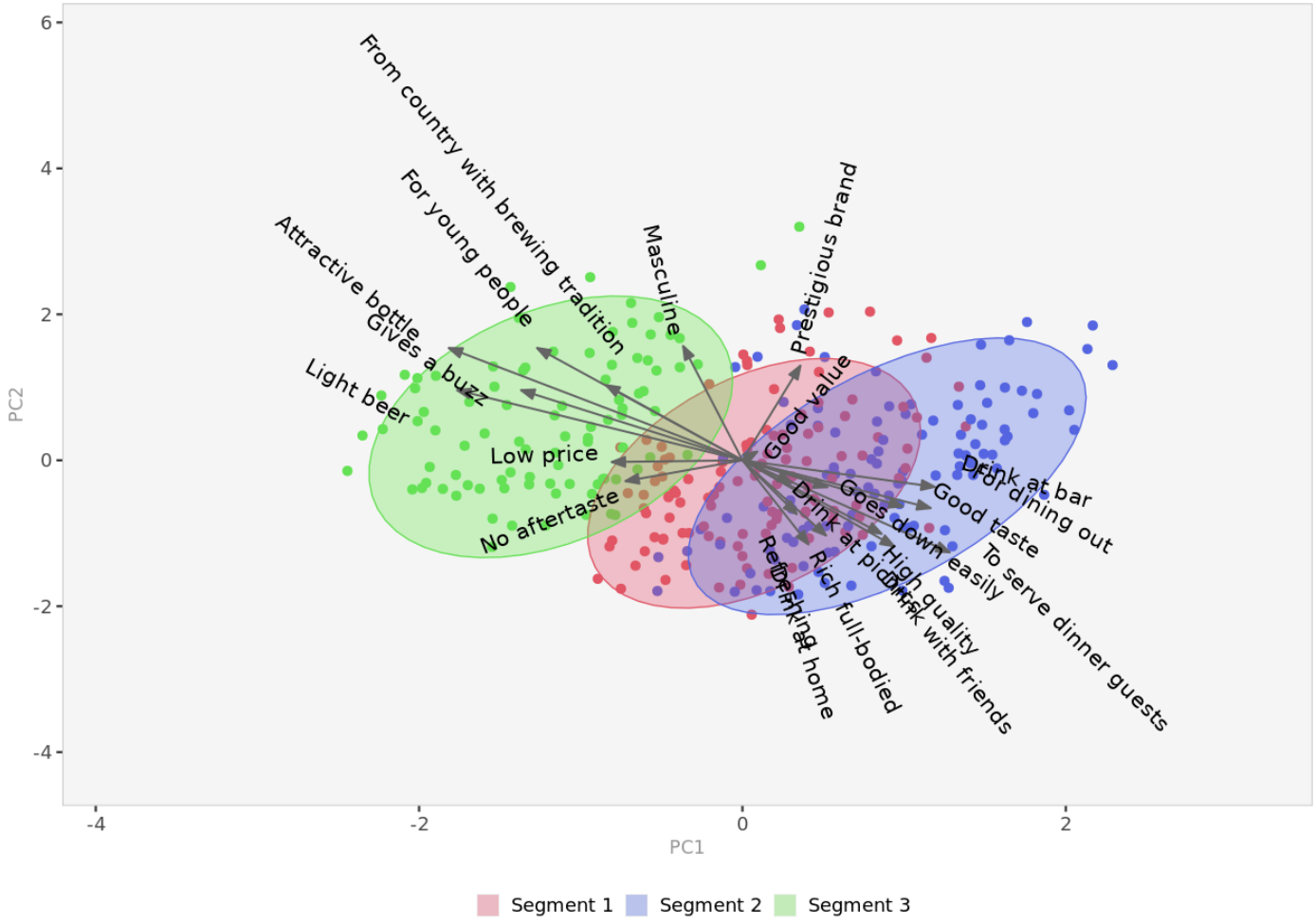
Segment differences per segment. Cell colors indicate to what extent a segment is statistically different from the rest of the population on each segmentation variable.

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 34.1% 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

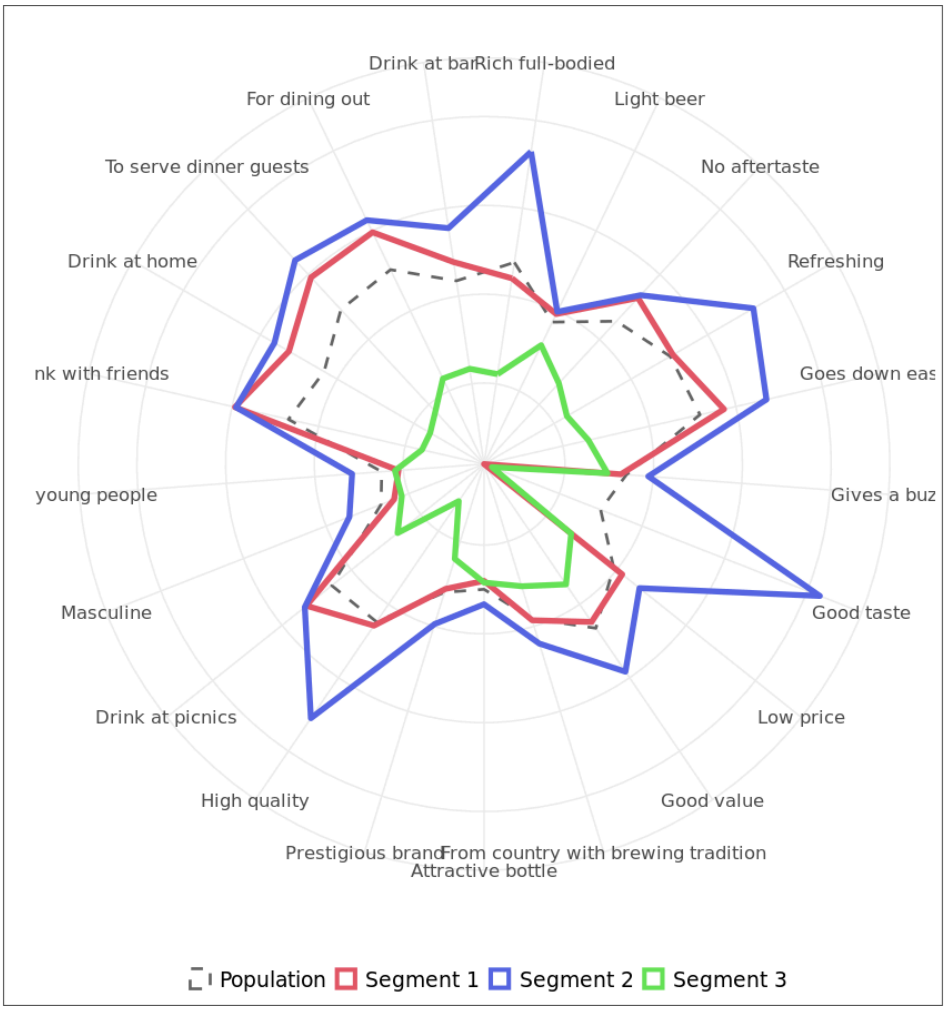
	Segment
6861	1
4129	1
4393	2
445	3
7393	2
964	1
6773	2
461	1
7156	2
5785	2

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.

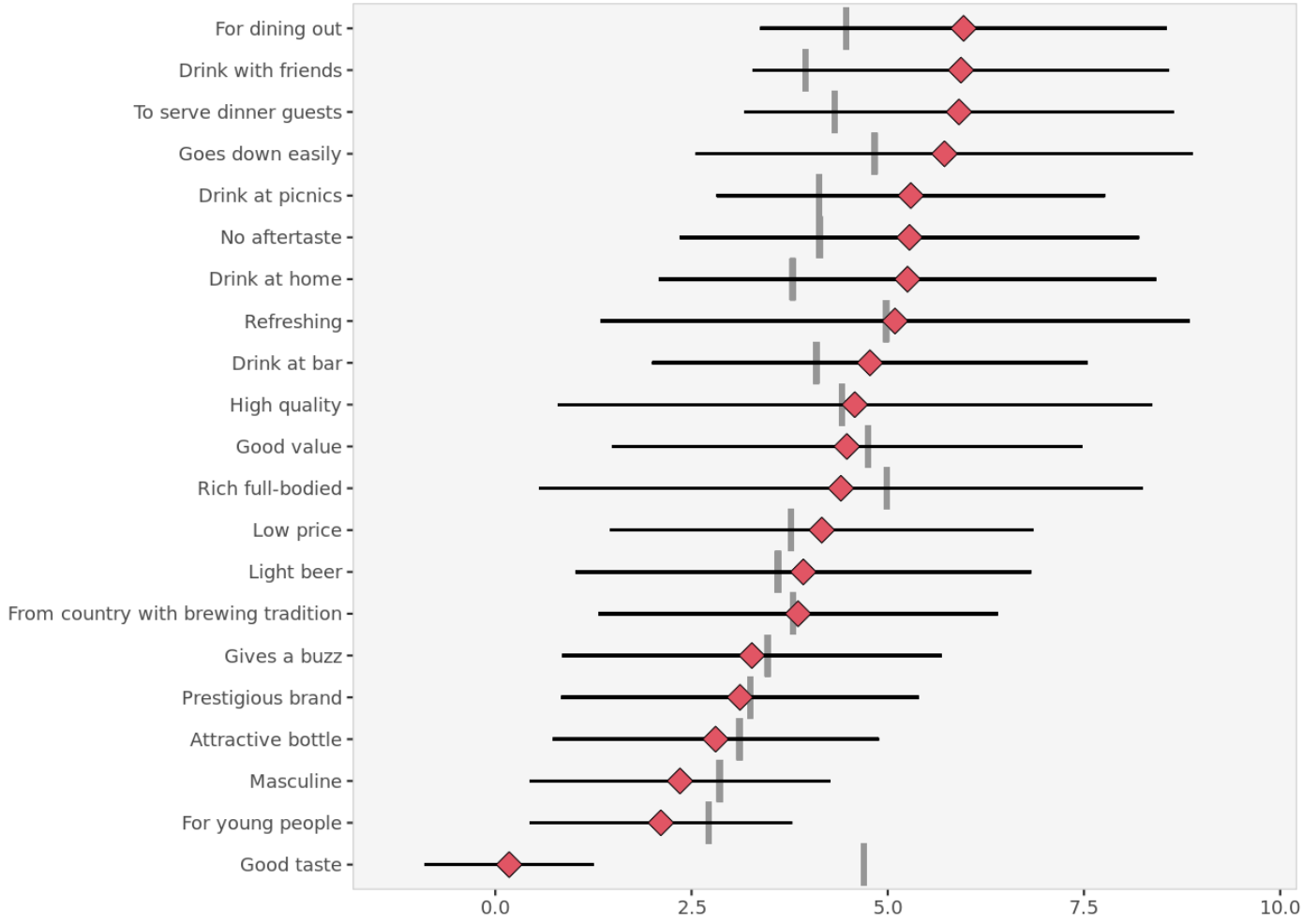


Spider chart.

Segment 1 profile

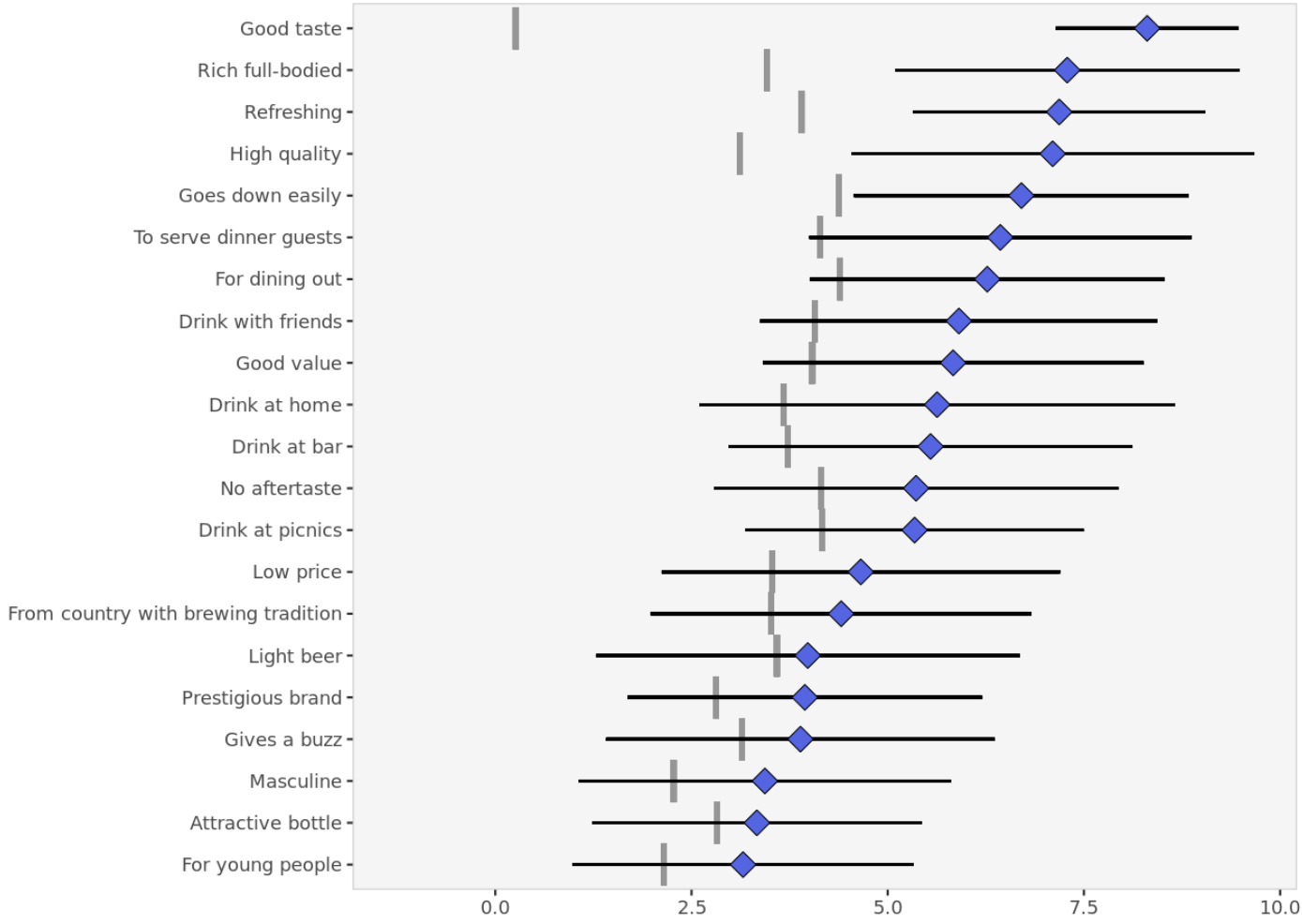
The following charts represent the profile of each segment. These charts are only available when the data are not standardized, hence the model assumes that all segmentation variables use the same scale.

- 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.



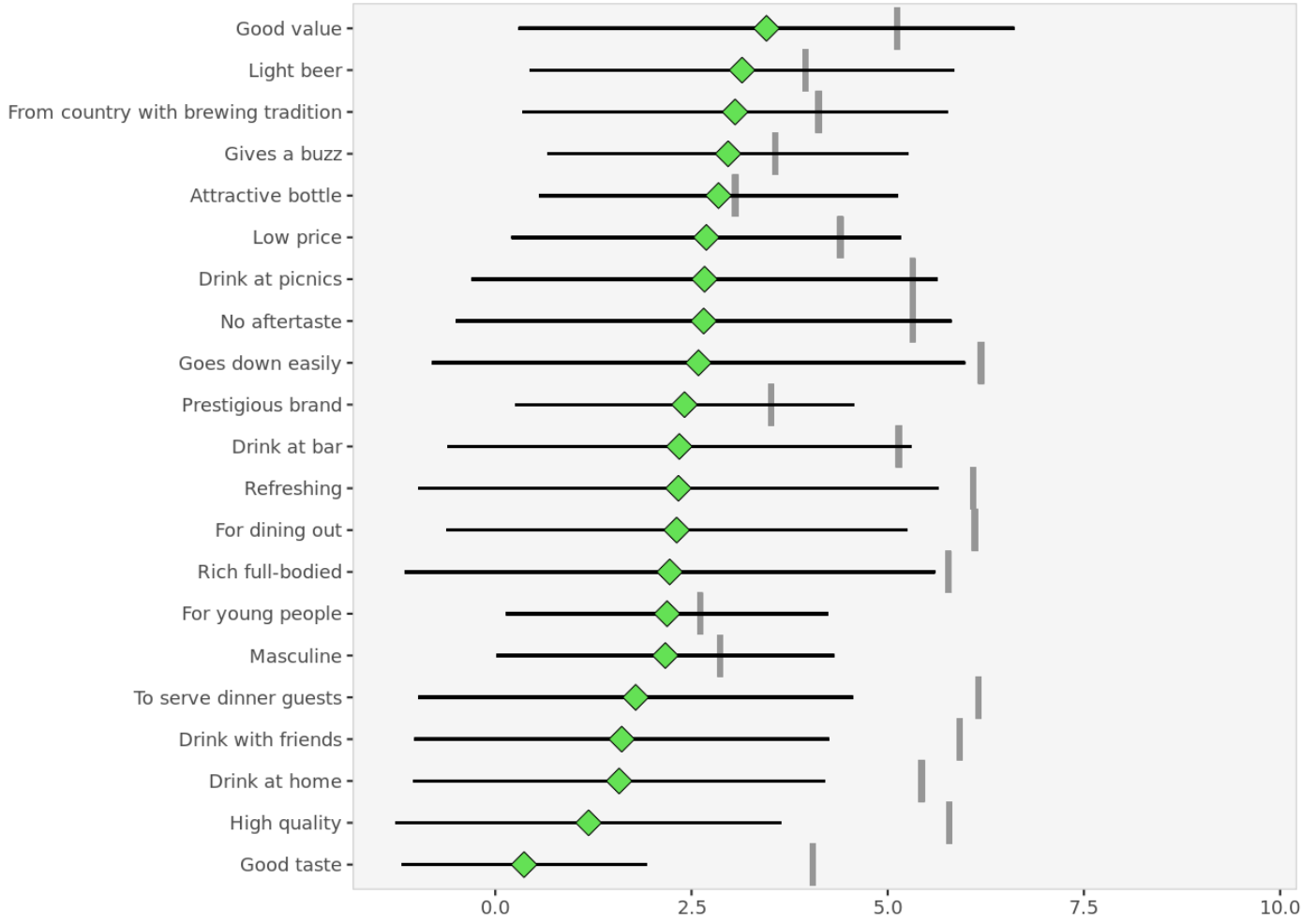
Segment 1 profile.

Segment 2 profile



Segment 2 profile.

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
`Weekly consumption`	9.42	9.10	9.81	9.38
`Age (1-7)`	4.77	4.79	4.68	4.86
`Income (1-7)`	5.45	5.47	5.39	5.50
`Education (1-6)`	4.47	4.49	4.65	4.24
`Sex (male=1)`	1.12	1.13	1.08	1.14
`Adapt to new situations`	3.50	3.53	3.45	3.52
`Make friends easily`	3.31	3.39	3.23	3.30
`Do not like to be tied to timetable`	3.56	3.55	3.57	3.56
`Like to take chances`	3.13	3.10	3.09	3.22
`Like to travel abroad`	3.31	3.39	3.23	3.29
`Like ethnic food`	3.47	3.51	3.39	3.51
`Knowledgeable about beer`	2.95	2.96	2.98	2.90

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.

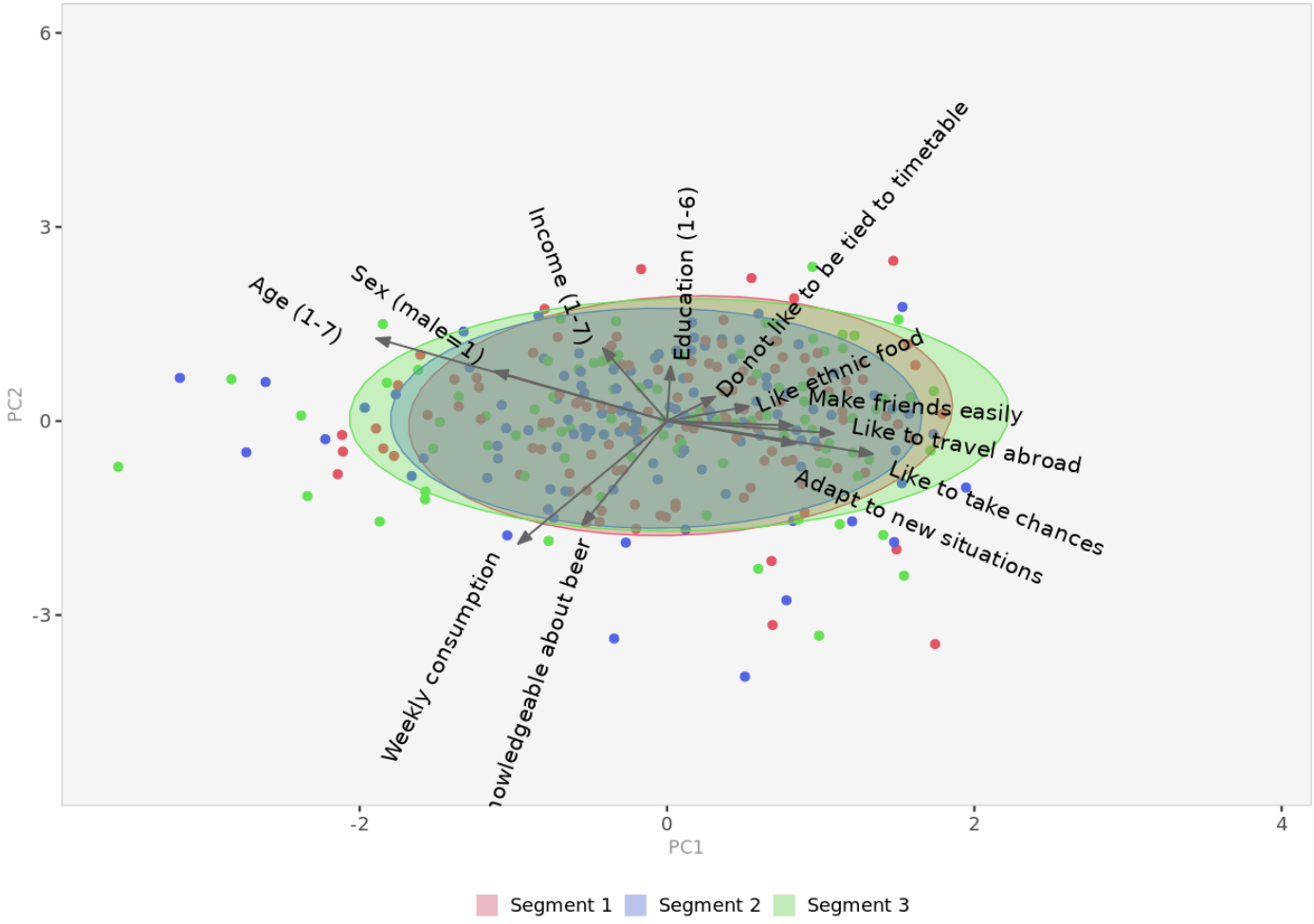
Descriptor space

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 30.3% 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)	1.872	-0.156
Weekly consumption	0.007	0.001
Age (1-7)	-0.052	0.044
Income (1-7)	-0.054	0.029
Education (1-6)	0.149	-0.135
Sex (male=1)	-0.464	0.204
Adapt to new situations	-0.166	0.028
Make friends easily	-0.200	-0.233
Do not like to be tied to timetable	0.138	-0.046
Like to take chances	0.070	0.369
Like to travel abroad	-0.179	-0.154
Like ethnic food	-0.165	0.061
Knowledgeable about beer	0.013	-0.057

Model parameters. Segment 1 is the model baseline.

P-values

	Segment 2	Segment 3
(Intercept)	0.223	0.923
Weekly consumption	0.657	0.965
Age (1-7)	0.621	0.684
Income (1-7)	0.568	0.770
Education (1-6)	0.165	0.200
Sex (male=1)	0.312	0.632
Adapt to new situations	0.489	0.915
Make friends easily	0.290	0.247
Do not like to be tied to timetable	0.552	0.845
Like to take chances	0.731	0.090

	Segmentation	
Like to travel abroad	0.286	0.385
Like ethnic food	0.408	0.777
Knowledgeable about beer	0.938	0.740

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

Confusion matrix

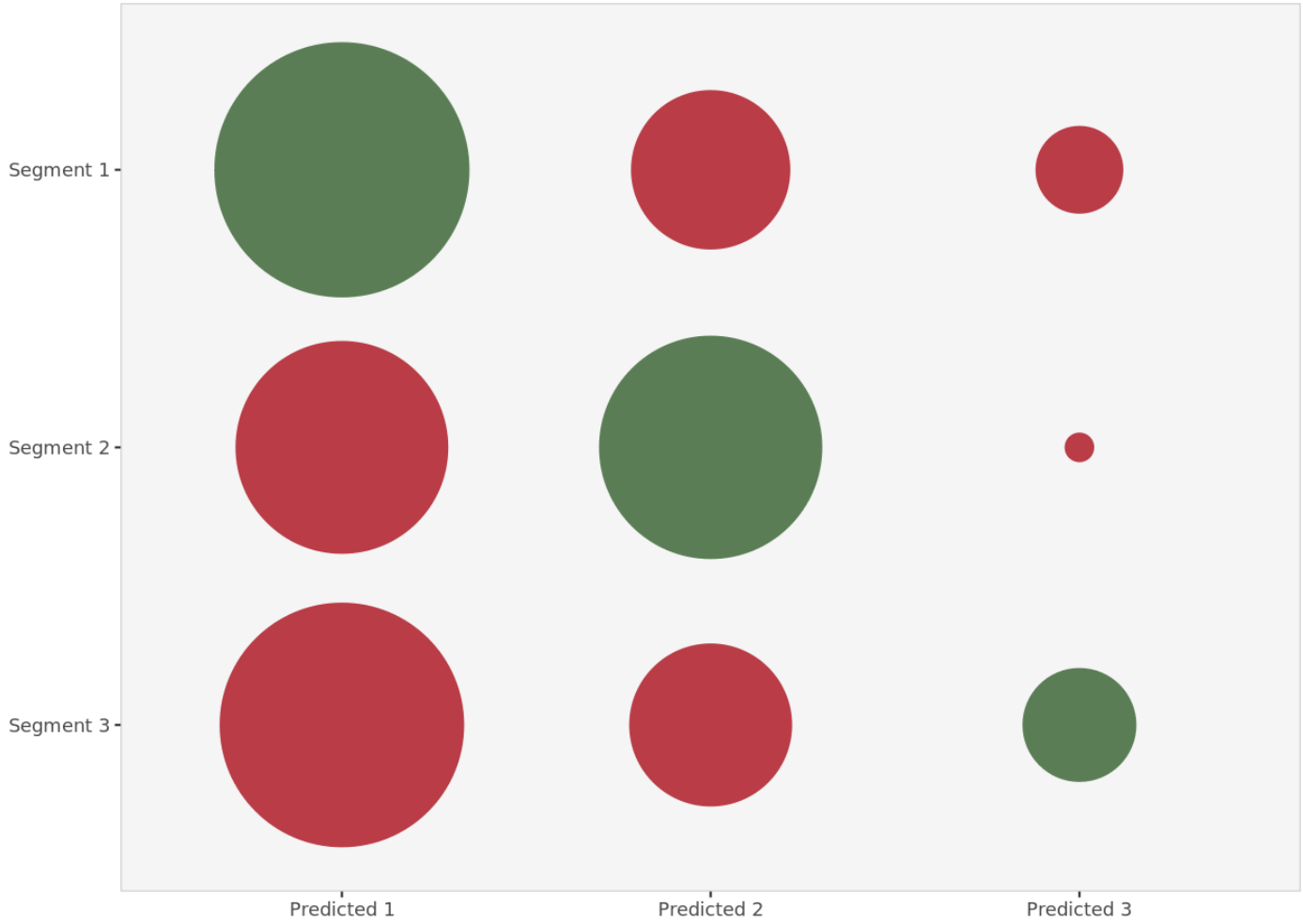
The confusion matrix compares actual segment membership (obtained from the segmentation analysis and the original segmentation variables) and predicted segment membership (obtained from the in-sample classification analysis and the descriptors alone). When actual and predicted segment memberships coincide, 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	70	32	17	119
Segment 2	46	50	12	108
Segment 3	49	25	16	90
Total	165	107	45	317

Confusion matrix (count). The model has correctly classified 136 of the 317 observations. The off-diagonal elements are classification errors.

	Predicted 1	Predicted 2	Predicted 3	Total
Segment 1	59%	27%	14%	100%
Segment 2	43%	46%	11%	100%
Segment 3	54%	28%	18%	100%

Confusion matrix (%). The global hit rate of the model is 43%. 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
6861	40%	35%	25%	1	1	1
4129	36%	18%	46%	3	1	0
4393	41%	35%	24%	1	2	0
445	35%	33%	32%	1	3	0
7393	36%	41%	24%	2	2	1
964	43%	17%	40%	1	1	1
6773	24%	55%	21%	2	2	1
461	39%	35%	26%	1	1	1
7156	35%	31%	34%	1	2	0
5785	35%	36%	29%	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.