enginius
MARKETING ENGINEERING ONLINE

Enginius

Segmentation

Anandita Maurya, Northeastern University

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.

The number of observations is too large to perform hierarchical clustering. Using Kmeans instead.

Data transformation

Standardization has not been performed.

Segment solution

4-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.

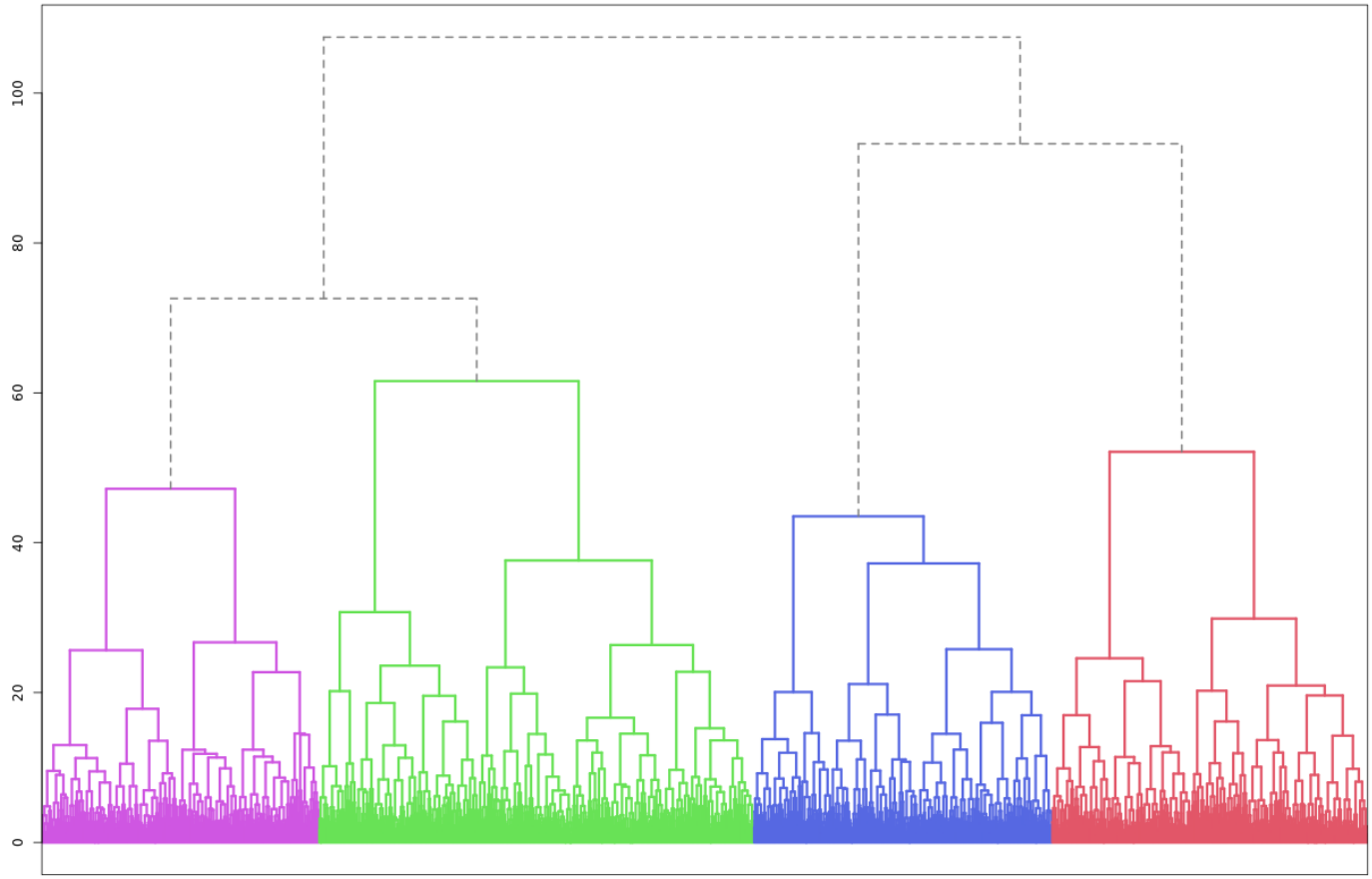
Using a statistical criteria exclusively (see scree plot analysis below), we have retained 4 segments.

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.



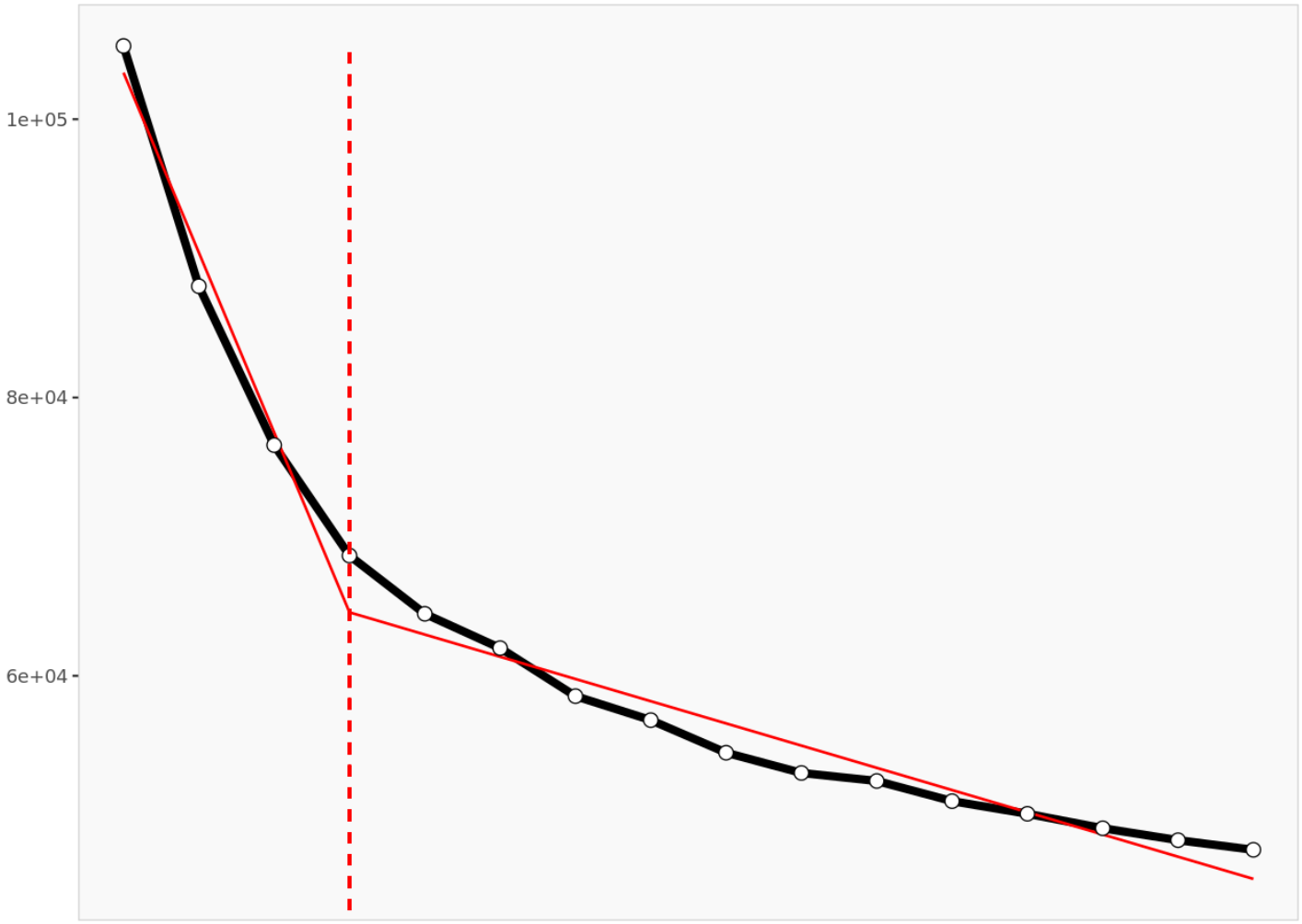
Dendrogram. The dendrogram is a tree diagram to illustrate the arrangement of clusters produced by hierarchical clustering, and how the observations are incrementally clustered together.

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.

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	Segment 4
Size	4 999	1 507	1 044	1 309	1 139
Relative size	100%	30%	21%	26%	23%

Segment size.

Segment description

	Population	Segment 1	Segment 2	Segment 3	Segment 4
Departure.and.Arrival.Time.Convenience	2.88	2.58	2.37	4.28	2.16
Ease.of.Online.Booking	2.77	2.35	2.27	4.21	2.12
Check.in.Service	2.92	3.47	3.09	2.47	2.55
Online.Boarding	2.79	2.80	2.05	3.10	3.09
Gate.Location	2.98	2.72	2.70	4.25	2.12
On.board.Service	2.89	3.69	3.33	2.51	1.86
Seat.Comfort	3.01	3.79	1.71	3.02	3.15
Leg.Room.Service	2.86	3.33	3.30	2.79	1.93
Cleanliness	2.76	3.79	1.55	2.62	2.67
Food.and.Drink	2.92	3.86	1.56	2.87	2.98
In.flight.Service	3.21	4.09	3.84	2.77	1.97
In.flight.Wifi.Service	2.40	2.70	2.12	2.67	1.96
In.flight.Entertainment	2.71	3.99	1.83	2.57	1.99
Baggage.Handling	3.20	4.04	3.83	2.79	2.00

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).

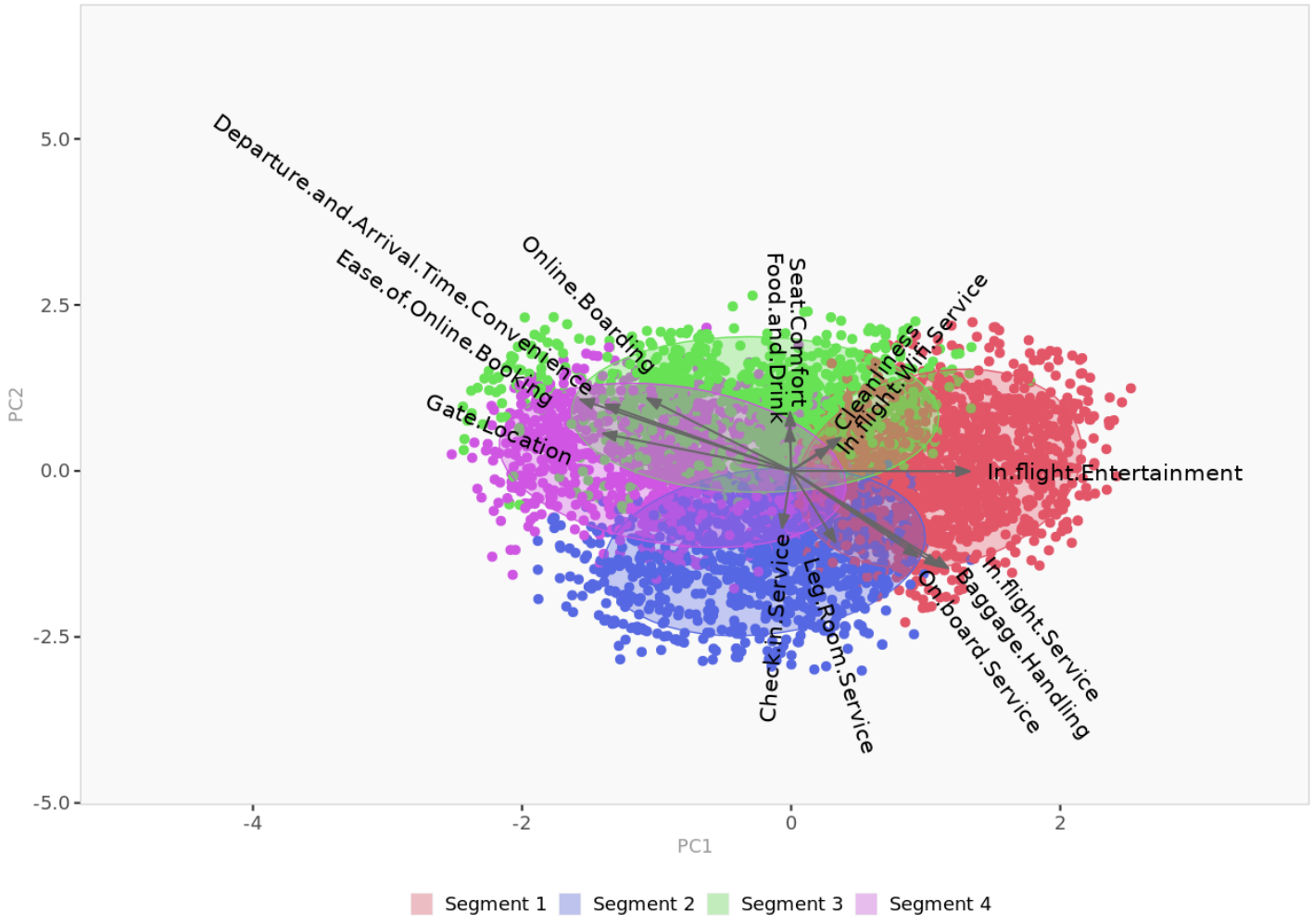
Segment differences per segment. Cell colors indicate to what extent a segment is statistically different from the rest of the population on each segmentation variable.

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 44.5% 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

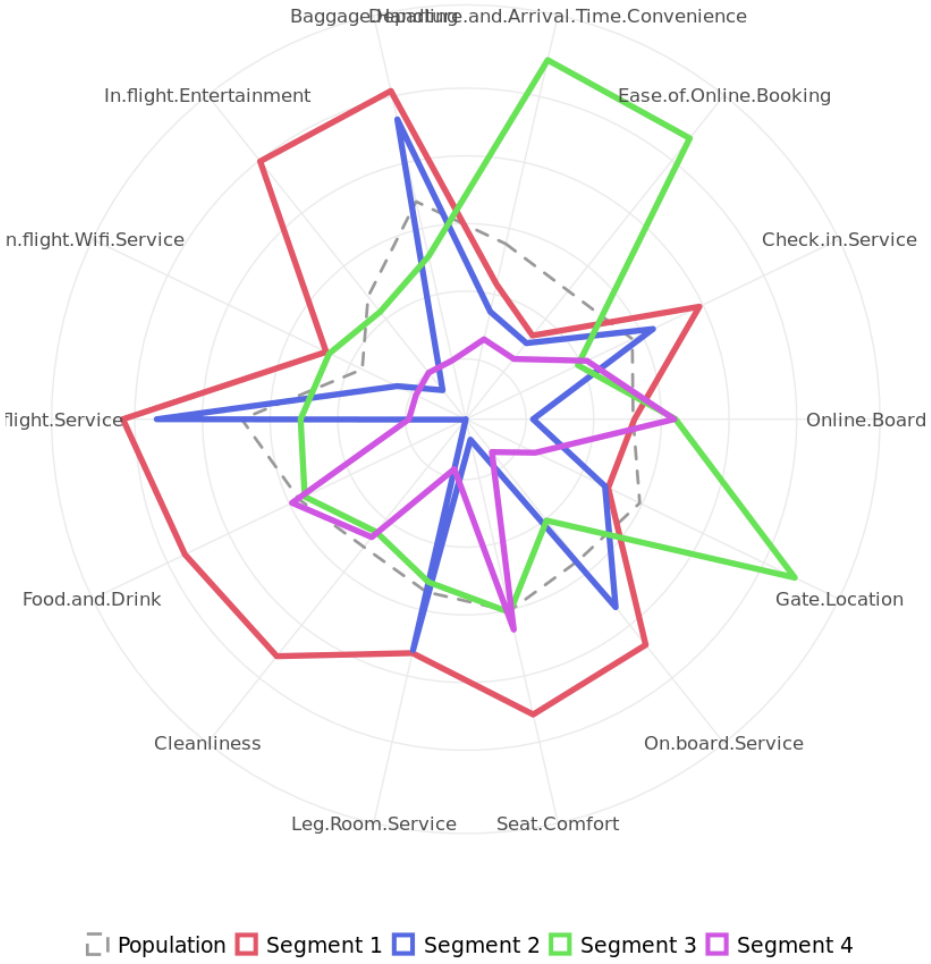
	Segment
43597	3
78707	4
93737	1
86157	1
124414	1
55900	1
89409	2
120054	2
14628	3
118322	2

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.



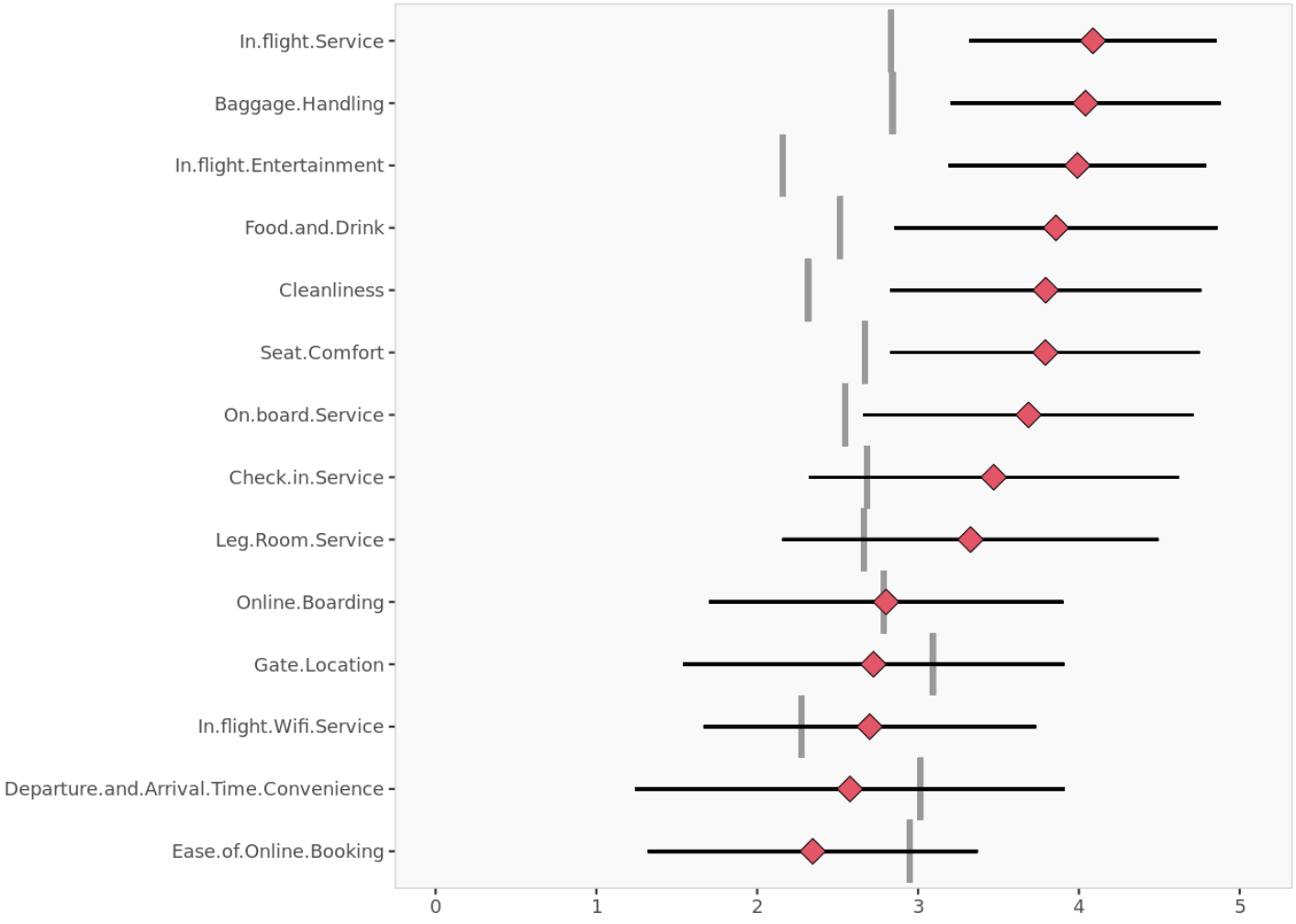
Spider chart.

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.

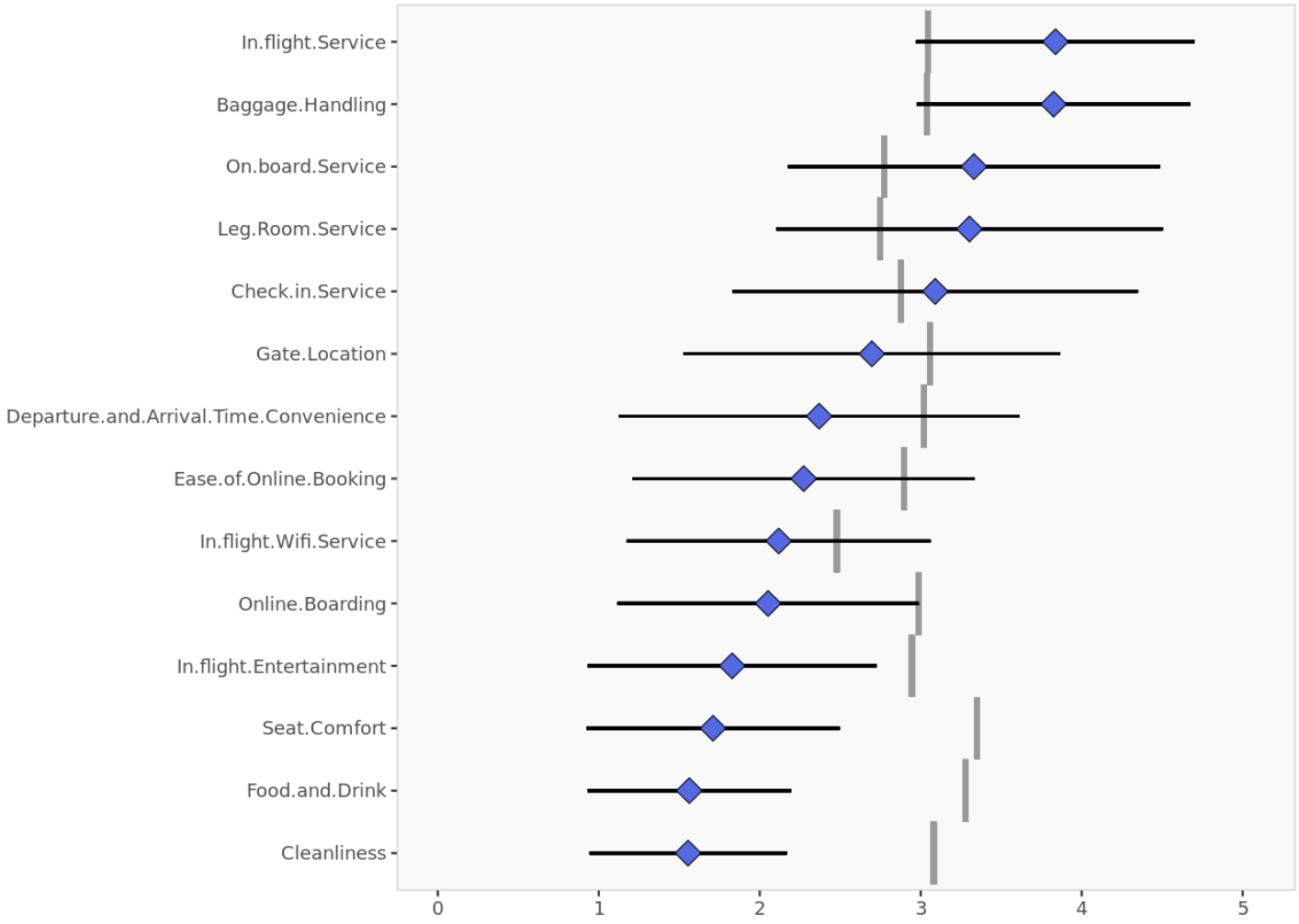
- For each segment, the segmentation variables are ordered in decreasing order of magnitude.
- The colored dots represent the average of the segment.
- The horizontal lines represent the standard deviations within that segment.

- The vertical, gray lines represent the averages of the rest of the population, after excluding members of the segment under scrutiny.



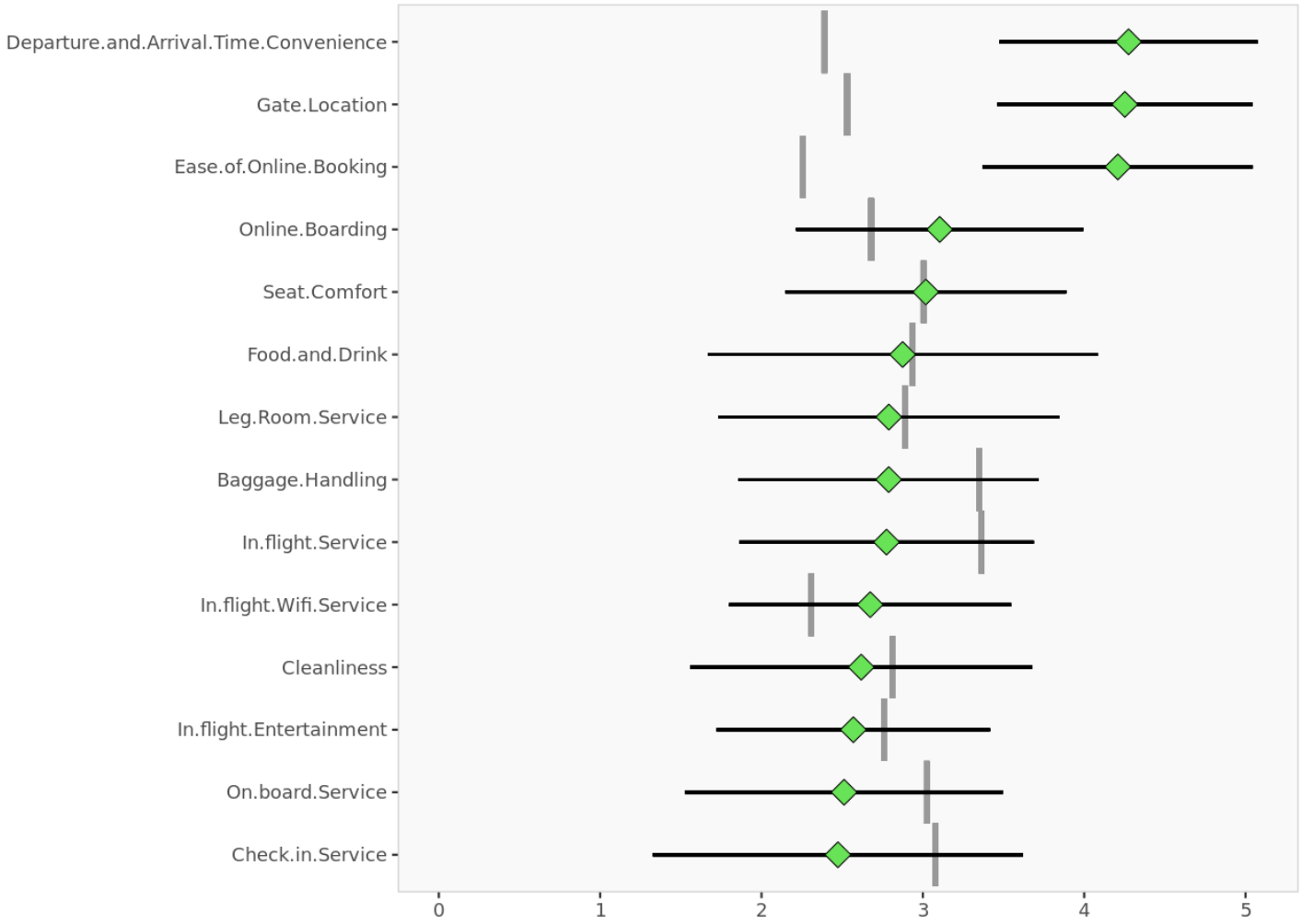
Segment 1 profile.

Segment 2 profile



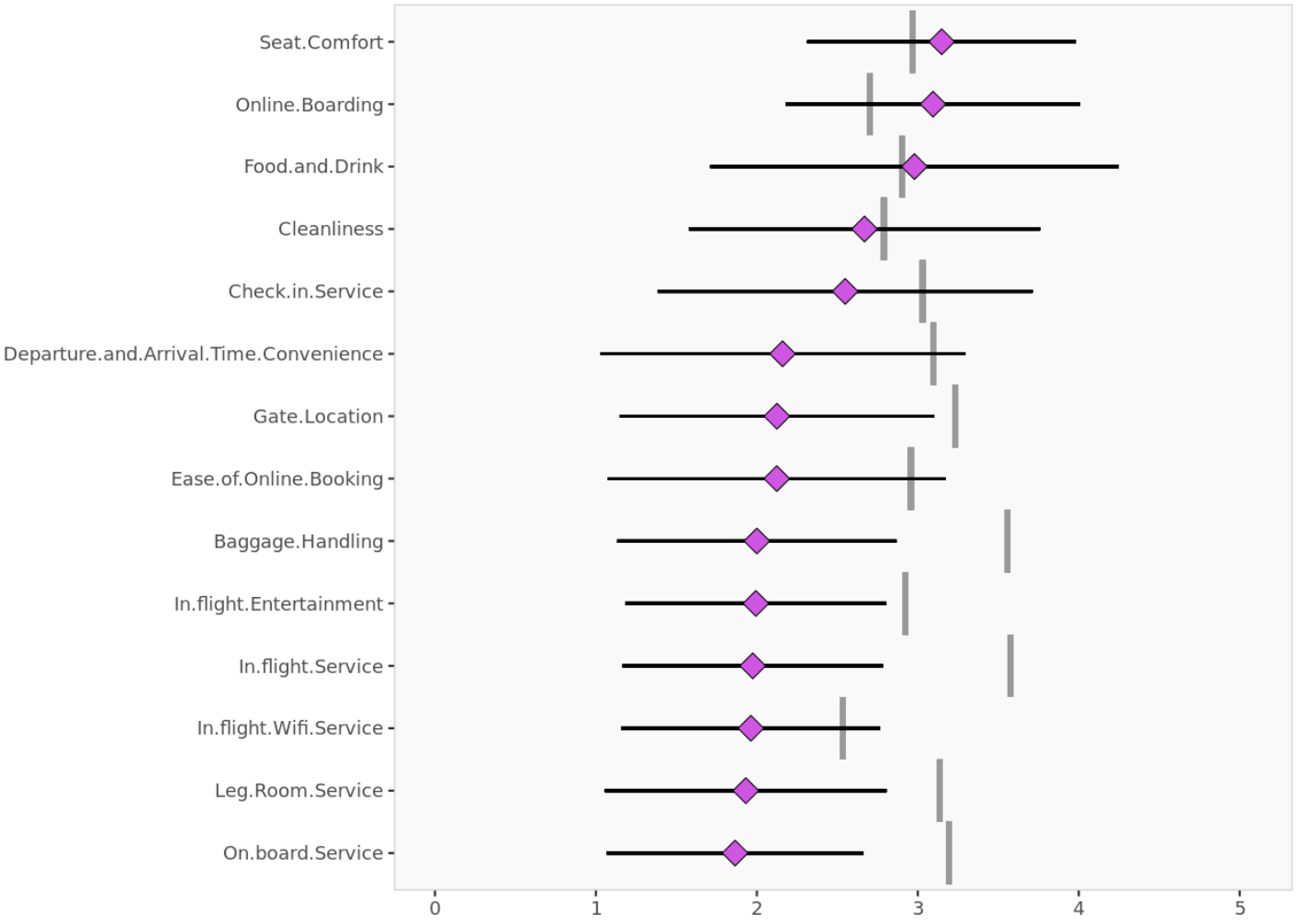
Segment 2 profile.

Segment 3 profile



Segment 3 profile.

Segment 4 profile



Segment 4 profile.

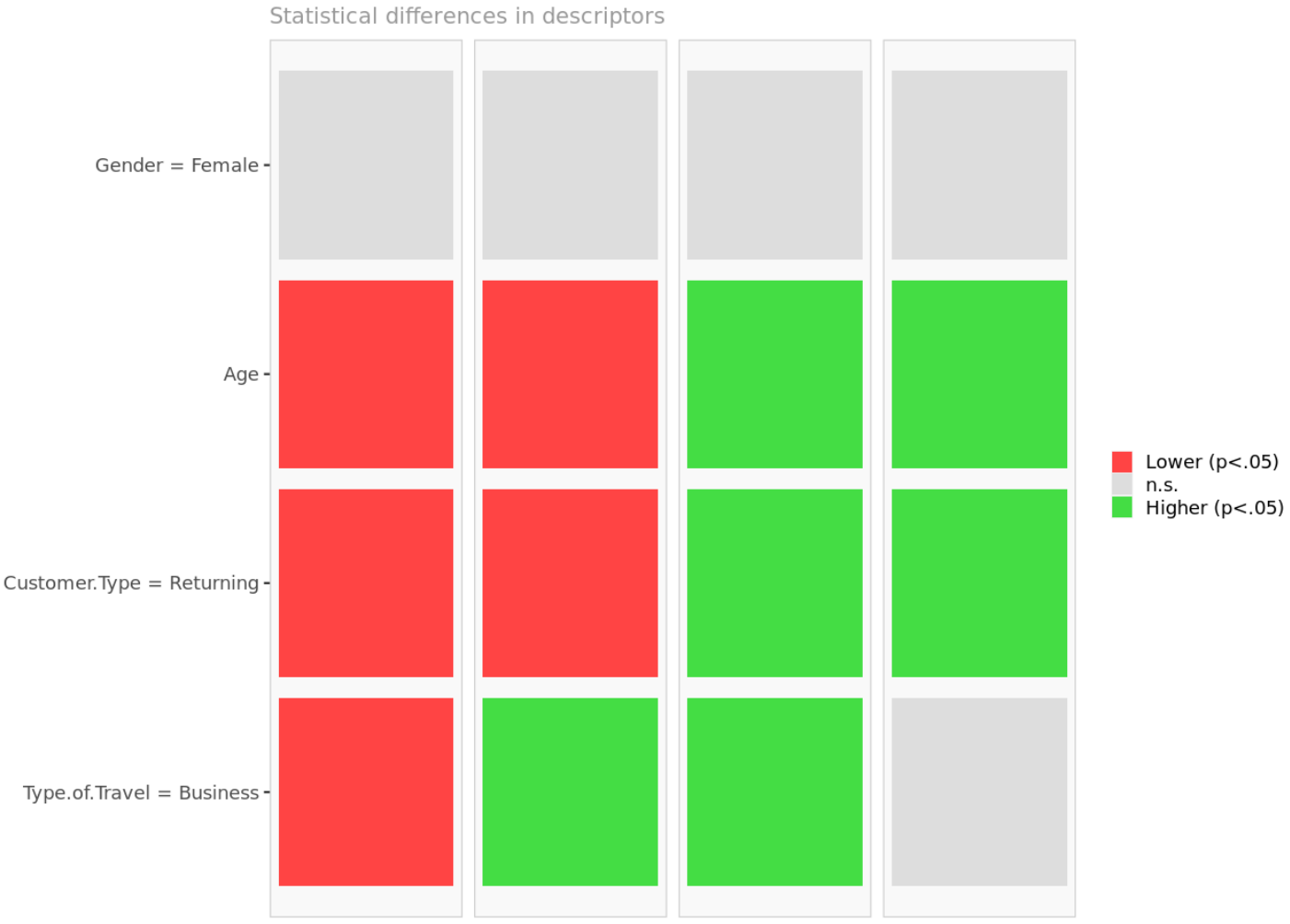
Descriptor analysis

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	Segment 4
Gender = Female	0.521	0.525	0.527	0.523	0.507
Age	39.3	37.1	33.6	41.5	44.8
Customer.Type = Returning	0.711	0.472	0.505	0.936	0.958
Type.of.Travel = Business	0.871	0.804	0.896	0.925	0.874

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.

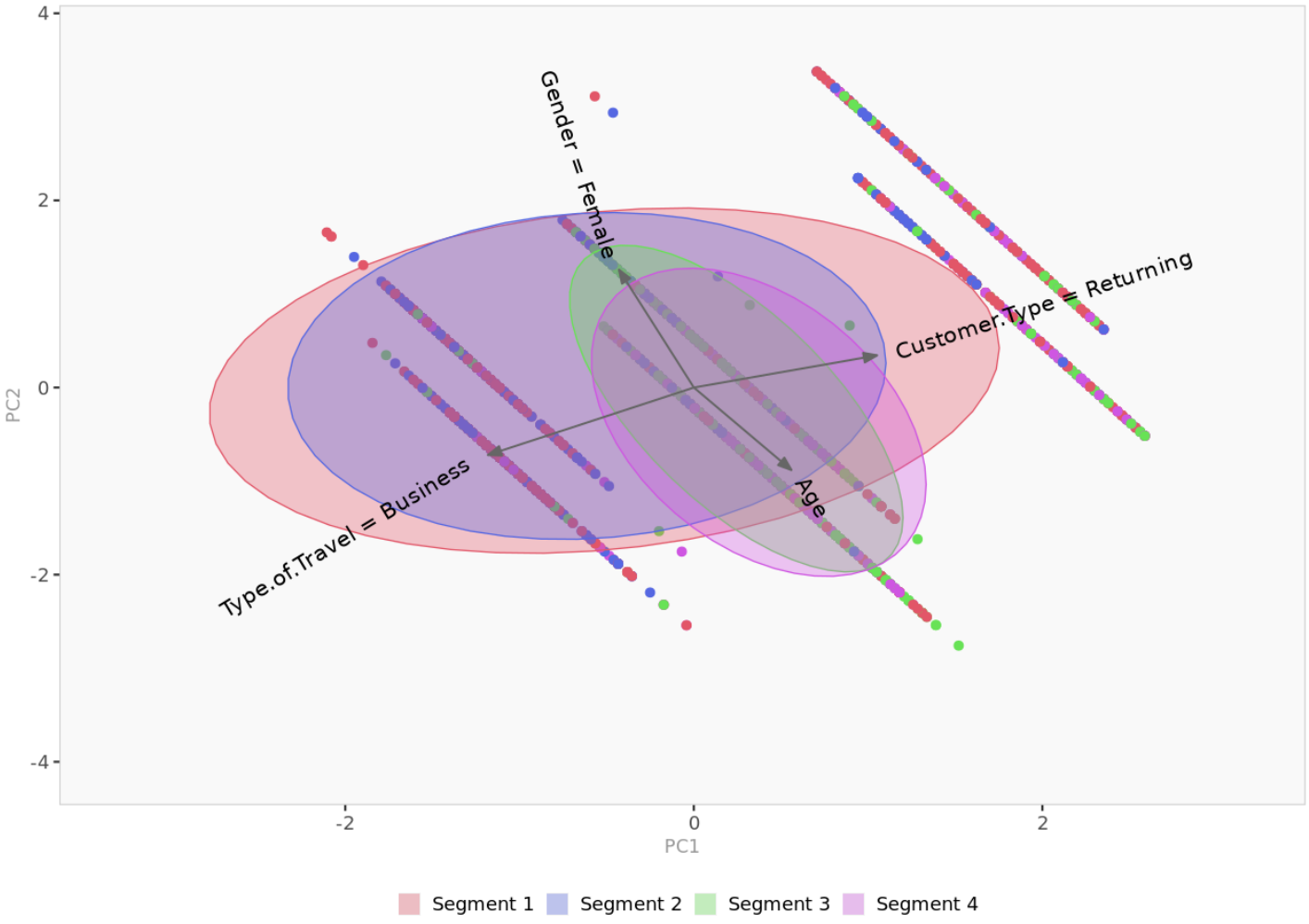
Descriptor space

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 57.7% 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	Segment 4
(Intercept)	-0.624	-4.869	-5.434
Gender = Female	-0.006	0.046	0.001
Age	-0.025	0.014	0.029
Customer.Type = Returning	0.472	3.181	3.496
Type.of.Travel = Business	1.06	2.07	1.55

Model parameters. Segment 1 is the model baseline.

P-values

	Segment 2	Segment 3	Segment 4
(Intercept)	0.001	0.000	0.000
Gender = Female	0.943	0.594	0.992
Age	0.000	0.000	0.000
Customer.Type = Returning	0.000	0.000	0.000
Type.of.Travel = Business	0.000	0.000	0.000

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

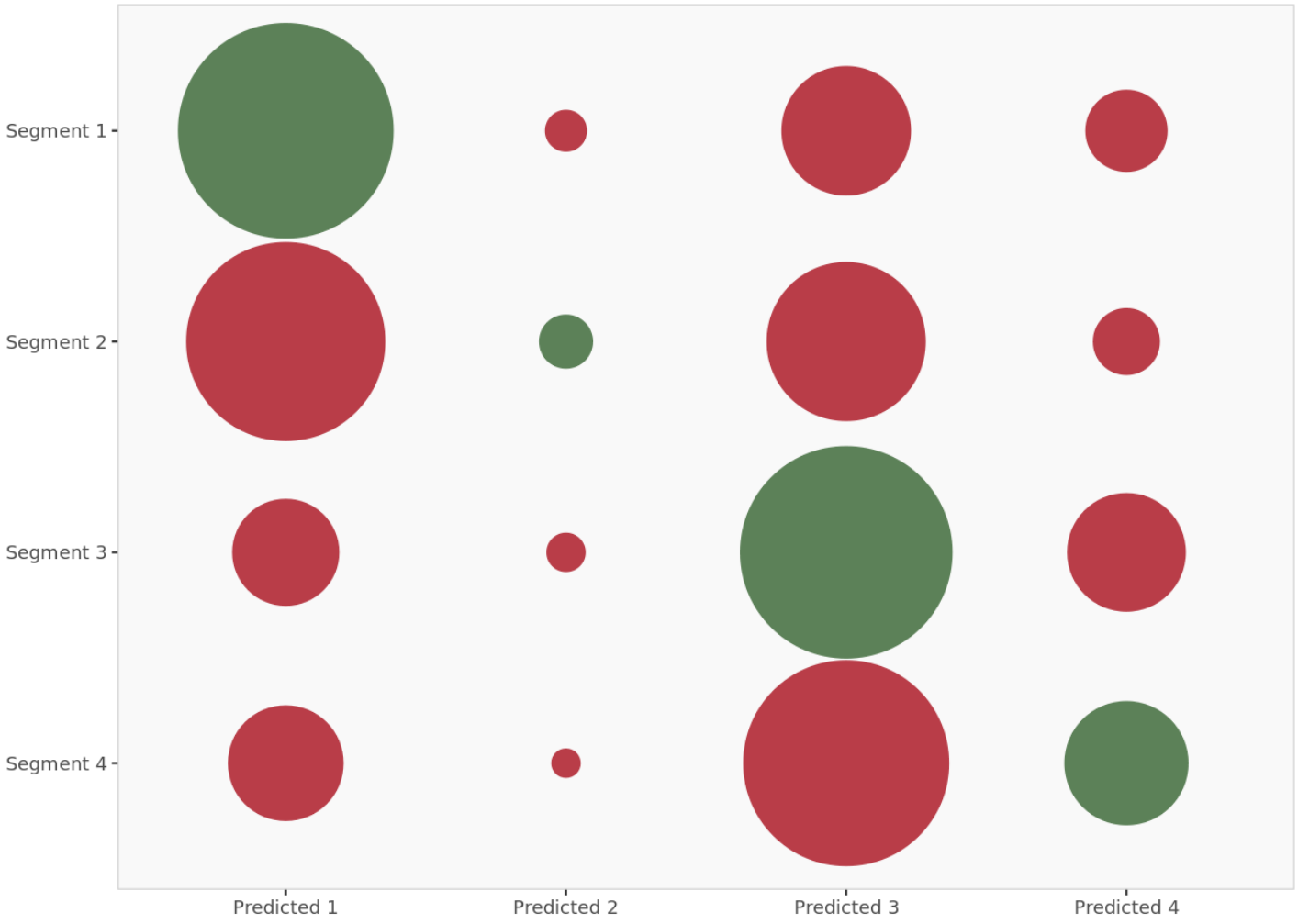
the in-sample classification analysis and the descriptors alone). When actual and predicted segment memberships coincide, the diagonal elements will be comparatively large, indicating that the classification model based on available descriptors is accurate.

	Predicted 1	Predicted 2	Predicted 3	Predicted 4	Total
Segment 1	1072	18	319	98	1507
Segment 2	618	22	365	39	1044
Segment 3	171	14	901	223	1309
Segment 4	182	10	730	217	1139
Total	2043	64	2315	577	4999

Confusion matrix (count). The model has correctly classified 2212 of the 4999 observations. The off-diagonal elements are classification errors.

	Predicted 1	Predicted 2	Predicted 3	Predicted 4	Total
Segment 1	71%	1%	21%	7%	100%
Segment 2	59%	2%	35%	4%	100%
Segment 3	13%	1%	69%	17%	100%
Segment 4	16%	1%	64%	19%	100%

Confusion matrix (%). The global hit rate of the model is 44%. 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)	Prob(cluster 4)	Predicted	Actual	Correct
43597	10%	4%	39%	46%	4	3	0
78707	15%	13%	40%	32%	3	4	0
93737	49%	30%	11%	10%	1	1	1
86157	56%	34%	6%	3%	1	1	1
124414	59%	29%	7%	5%	1	1	1
55900	18%	34%	32%	16%	2	1	0
89409	62%	23%	9%	7%	1	2	0
120054	51%	41%	5%	2%	1	2	0
14628	14%	12%	39%	34%	3	3	1
118322	51%	41%	5%	2%	1	2	0

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.