A Market Segmentation and Purchase Drivers Process

T. Evgeniou

IMPORTANT Please make sure you create a copy of this file with a customised name, so that your work (e.g. answers to the questions) is not over-written when you pull the latest content from the course github.

This is a **template process for market segmentation based on survey data**, using the Boats cases A and B.

The process is split in 3 parts, corresponding to Sessions 3-4, 5-6, and an optional last part: 1. *Part 1*: We use some of the survery questions (e.g. in this case the first 29 "attitude" questions) to construct key variables ("factors") using *dimensionality reduction* that will be used for the segmentation. 2. *Part 2*: We use the selected variables to segment the market using *cluster analysis* and profile the segments using profiling variables that we select. 3. *Part 3*: For the market segments we create, we will use *classification analysis* to classify people based on whether or not they have purchased a product and find what are the key purchase drivers per segment.

Finally, we will use the results of this analysis to make business decisions e.g. about brand positioning, product development, etc depending on our market segments and key purchase drivers we find at the end of this process.

Before starting, make sure you have pulled the course files on your github repository. In case you would like to manually go through the commands below, make sure you are in the directory of this file. For example, assuming we are now in the "MYDIRECTORY/INSEADAnalytics" directory, we can do these:

As always, you can use the `help` command in Rstudio to find out about any R function (e.g. type `help(list.files)` to learn what the R function `list.files` does).

Let's start.

```
## Creating a generic function for 'toJSON' from package 'jsonlite' in package 'googleVis'
```

The Data

First we load the data to use.

Part 1: Dimensionality Reduction

The code used here is along the lines of the code in the session 3-4 reading FactorAnalysisReading.Rmd. We follow the process described in the Dimensionality Reduction reading.

In this part we also become familiar with:

1. Some visualization tools;
2. Principal Component Analysis and Factor Analysis;
3. Introduction to machine learning methods;

All user inputs for this part should be selected here:

Steps 1/2: Check the Data

Start by some basic visual exploration of, say, a few data:

	Q1 1	Q1 2	Q1 3	Q1 4	Q1 5	Q1 6	Q1 7	Q1 8	Q1 9	Q1 10	Q1 11	Q1 12	Q
observation 01	5	1	5	5	5	5	5	3	5	4	2	1	
observation 02	3	2	4	4	4	4	5	3	4	4	3	2	
observation 03	3	1	4	4	5	4	4	2	4	3	2	2	
observation 04	5	2	3	4	5	5	3	3	3	4	4	2	
observation 05	4	2	4	5	5	4	5	4	5	4	5	1	
observation 06	4	2	2	4	5	5	4	3	2	4	2	2	
observation 07	4	3	4	4	5	5	5	5	3	1	5	3	
observation 08	4	3	4	4	4	5	4	4	5	3	4	1	
observation 09	4	3	4	4	4	4	4	3	4	3	5	1	
observation 10	2	5	3	2	1	3	5	4	3	4	1	3	

The data we use here have the following descriptive statistics:

	min	25 percent	median	mean	75 percent	max	std
Q1 1	1	4	4	4.03	5	5	0.82
Q1 2	1	2	3	2.89	4	5	1.01
Q1 3	1	2	3	3.12	4	5	1.02
Q1 4	1	3	4	3.89	4	5	0.82
Q1 5	1	3	4	3.55	4	5	0.93
Q1 6	1	4	4	3.95	4	5	0.82
Q1 7	1	3	4	3.67	4	5	0.90
Q1 8	1	3	4	3.74	4	5	0.82
Q1 9	1	2	3	2.89	4	5	1.08
Q1 10	1	3	3	3.37	4	5	0.93
Q1 11	1	3	4	3.46	4	5	1.15
Q1 12	1	2	3	2.86	4	5	1.01
Q1 13	1	2	3	3.02	4	5	0.98
Q1 14	1	3	3	3.25	4	5	0.97
Q1 15	1	3	4	3.63	4	5	0.89
Q1 16	1	2	3	3.10	4	5	1.05
Q1 17	1	2	3	3.08	4	5	0.98
Q1 18	1	4	4	4.12	5	5	0.74
Q1 19	1	4	4	4.20	5	5	0.72
Q1 20	1	2	3	3.16	4	5	0.97
Q1 21	1	4	4	4.25	5	5	0.73
Q1 22	1	4	4	4.01	4	5	0.74
Q1 23	1	3	4	3.56	4	5	1.02
Q1 24	1	4	4	4.11	5	5	0.76
Q1 25	1	3	4	3.79	4	5	0.91
Q1 26	1	2	3	2.95	4	5	1.05
Q1 27	1	2	3	3.16	4	5	1.05
Q1 28	1	3	3	3.31	4	5	0.98
Q1 29	1	4	4	4.03	4	5	0.73

Step 3: Check Correlations

This is the correlation matrix of the customer responses to the 29 attitude questions - which are the only questions that we will use for the segmentation (see the case):

	Q1 1	Q1 2	Q1 3	Q1 4	Q1 5	Q1 6	Q1 7	Q1 8	Q1 9	Q1 10	Q1 11	Q1 12	Q1 13	Q1 14	Q1 1
Q1 1	1.00	0.01	0.11	0.20	0.18	0.27	0.18	0.09	0.08	0.11	0.14	-0.05	0.12	0.18	0.2
Q1 2	0.01	1.00	-0.03	-0.21	-0.21	-0.04	0.02	0.20	0.09	0.16	0.04	0.37	0.01	-0.03	-0.0
Q1 3	0.11	-0.03	1.00	0.26	0.40	0.34	0.44	-0.05	0.58	0.14	0.10	-0.09	0.48	0.46	0.3
Q1 4	0.20	-0.21	0.26	1.00	0.37	0.20	0.18	0.00	0.17	0.10	0.06	-0.16	0.27	0.29	0.3
Q1 5	0.18	-0.21	0.40	0.37	1.00	0.29	0.29	-0.03	0.33	0.14	0.07	-0.17	0.45	0.46	0.4
Q1 6	0.27	-0.04	0.34	0.20	0.29	1.00	0.55	0.04	0.35	0.12	0.15	-0.12	0.29	0.31	0.3
Q1 7	0.18	0.02	0.44	0.18	0.29	0.55	1.00	-0.01	0.49	0.12	0.12	-0.11	0.35	0.36	0.3
Q1 8	0.09	0.20	-0.05	0.00	-0.03	0.04	-0.01	1.00	-0.09	0.09	0.14	0.24	-0.05	-0.02	0.0
Q1 9	0.08	0.09	0.58	0.17	0.33	0.35	0.49	-0.09	1.00	0.14	0.06	-0.04	0.48	0.43	0.3
Q1 10	0.11	0.16	0.14	0.10	0.14	0.12	0.12	0.09	0.14	1.00	-0.09	0.12	0.16	0.11	0.1
Q1 11	0.14	0.04	0.10	0.06	0.07	0.15	0.12	0.14	0.06	-0.09	1.00	0.09	0.08	0.13	0.2
Q1 12	-0.05	0.37	-0.09	-0.16	-0.17	-0.12	-0.11	0.24	-0.04	0.12	0.09	1.00	-0.11	-0.17	-0.1
Q1 13	0.12	0.01	0.48	0.27	0.45	0.29	0.35	-0.05	0.48	0.16	0.08	-0.11	1.00	0.64	0.4
Q1 14	0.18	-0.03	0.46	0.29	0.46	0.31	0.36	-0.02	0.43	0.11	0.13	-0.17	0.64	1.00	0.5
Q1 15	0.26	-0.08	0.38	0.30	0.42	0.31	0.34	0.06	0.33	0.11	0.20	-0.17	0.46	0.50	1.0
Q1 16	0.16	-0.02	0.39	0.18	0.36	0.27	0.31	0.02	0.39	-0.03	0.32	-0.02	0.43	0.43	0.4
Q1 17	0.15	0.04	0.38	0.17	0.32	0.24	0.29	0.05	0.37	-0.03	0.31	0.02	0.43	0.40	0.3
Q1 18	0.25	-0.04	0.24	0.18	0.23	0.44	0.40	0.07	0.22	0.14	0.11	-0.12	0.20	0.25	0.3
Q1 19	0.27	-0.04	0.14	0.19	0.18	0.36	0.28	0.09	0.07	0.09	0.12	-0.09	0.11	0.18	0.2
Q1 20	0.19	0.05	0.39	0.18	0.32	0.35	0.36	0.04	0.37	0.10	0.25	0.01	0.39	0.41	0.4
Q1 21	0.24	-0.10	0.18	0.18	0.19	0.42	0.33	0.06	0.14	0.08	0.13	-0.17	0.14	0.21	0.2
Q1 22	0.23	-0.08	0.28	0.23	0.27	0.41	0.39	0.05	0.23	0.09	0.17	-0.11	0.23	0.29	0.3
Q1 23	0.19	0.00	0.34	0.16	0.29	0.32	0.30	0.10	0.29	0.07	0.19	-0.03	0.32	0.36	0.3
Q1 24	0.21	-0.08	0.23	0.23	0.25	0.37	0.33	0.02	0.23	0.13	0.08	-0.17	0.20	0.21	0.2
Q1 25	0.23	0.01	0.36	0.22	0.29	0.42	0.42	0.10	0.32	0.08	0.25	-0.05	0.32	0.35	0.4
Q1 26	0.10	0.07	0.47	0.19	0.34	0.31	0.39	-0.04	0.50	0.13	0.09	-0.06	0.48	0.46	0.3
Q1 27	0.13	0.05	0.40	0.17	0.29	0.34	0.37	0.03	0.40	0.08	0.16	0.00	0.40	0.39	0.3
Q1 28	0.18	0.02	0.43	0.21	0.33	0.39	0.40	0.05	0.40	0.07	0.18	-0.01	0.40	0.40	0.3
Q1 29	0.20	-0.03	0.17	0.19	0.18	0.27	0.24	0.10	0.11	0.05	0.17	-0.04	0.19	0.21	0.2

Questions

1. Do you see any high correlations between the responses? Do they make sense?
2. What do these correlations imply?

Answers:

Step 4: Choose number of factors

Clearly the survey asked many redundant questions (can you think some reasons why?), so we may be able to actually "group" these 29 attitude questions into only a few "key factors". This not only will simplify the data, but will also greatly facilitate our understanding of the customers.

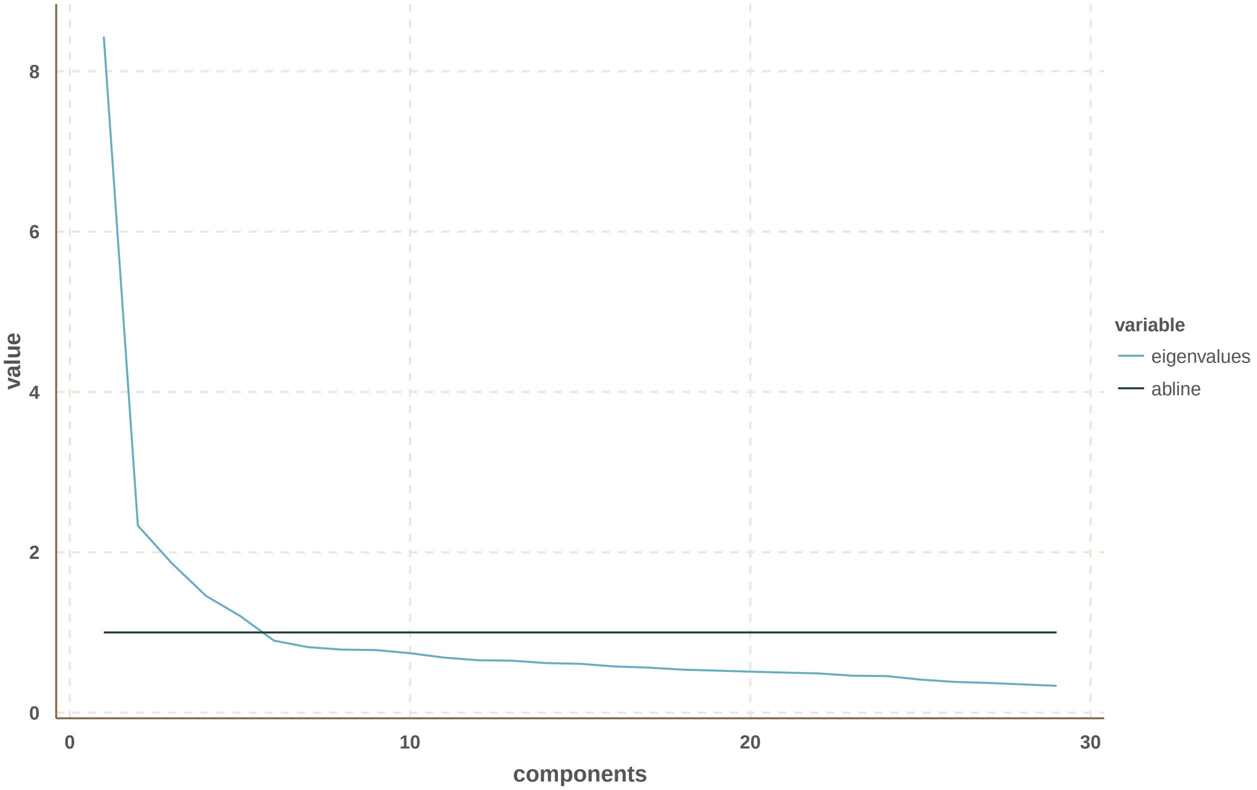
To do so, we use methods called Principal Component Analysis and factor analysis as also discussed in the Dimensionality Reduction readings. We can use two different R commands for this (they make slightly different information easily available as output): the command `principal` (check `help(principal)` from R package psych), and the command `PCA` from R package FactoMineR - there are more packages and commands for this, as these methods are very widely used.

Here is how the `principal` function is used:

Here is how we use the PCA function:

Let's look at the **variance explained** as well as the **eigenvalues** (see session readings):

	Eigenvalue	Pct of explained variance	Cumulative pct of explained variance
Component 1	8.43	29.08	29.08
Component 2	2.33	8.05	37.12
Component 3	1.86	6.42	43.55
Component 4	1.46	5.03	48.57
Component 5	1.21	4.16	52.74
Component 6	0.90	3.10	55.84
Component 7	0.82	2.82	58.65
Component 8	0.79	2.71	61.36
Component 9	0.78	2.69	64.05
Component 10	0.74	2.56	66.61
Component 11	0.69	2.37	68.98
Component 12	0.65	2.25	71.23
Component 13	0.65	2.23	73.47
Component 14	0.62	2.13	75.60
Component 15	0.61	2.10	77.70
Component 16	0.58	1.99	79.69
Component 17	0.56	1.94	81.62
Component 18	0.54	1.85	83.47
Component 19	0.52	1.81	85.28
Component 20	0.51	1.76	87.04
Component 21	0.50	1.72	88.77
Component 22	0.49	1.69	90.45
Component 23	0.46	1.59	92.04
Component 24	0.46	1.57	93.61
Component 25	0.41	1.42	95.03
Component 26	0.38	1.32	96.36
Component 27	0.37	1.28	97.63
Component 28	0.35	1.22	98.85
Component 29	0.33	1.15	100.00



Questions:

1. Can you explain what this table and the plot are? What do they indicate? What can we learn from these?
2. Why does the plot have this specific shape? Could the plotted line be increasing?
3. What characteristics of these results would we prefer to see? Why?

Your Answers here:

Step 5: Interpret the factors

Let's now see how the "top factors" look like.

To better visualise them, we will use what is called a "rotation". There are many rotations methods. In this case we selected the varimax rotation. For our data, the 5 selected factors look as follows after this rotation:

	Component 1	Component 2	Component 3	Component 4	Component 5
Q1 9	0.78	0.12	0.00	-0.12	-0.01
Q1 26	0.72	0.11	0.10	0.01	0.02
Q1 3	0.71	0.15	0.17	-0.04	-0.06
Q1 13	0.68	0.02	0.40	0.01	-0.03
Q1 27	0.63	0.24	-0.02	0.28	0.05
Q1 28	0.62	0.30	0.03	0.31	0.04
Q1 14	0.61	0.11	0.44	0.09	-0.07
Q1 16	0.57	0.09	0.14	0.55	-0.03
Q1 7	0.56	0.50	-0.05	-0.07	-0.03
Q1 17	0.55	0.04	0.15	0.54	0.03
Q1 20	0.55	0.25	0.11	0.36	0.10

	Component 1	Component 2	Component 3	Component 4	Component 5
Q1 5	0.42	0.14	0.58	0.03	-0.18
Q1 15	0.42	0.24	0.50	0.22	-0.03
Q1 23	0.41	0.29	0.15	0.31	0.07
Q1 25	0.39	0.43	0.15	0.25	0.06
Q1 6	0.38	0.62	0.02	0.00	-0.02
Q1 22	0.24	0.62	0.11	0.18	-0.05
Q1 18	0.18	0.73	0.09	0.00	0.01
Q1 10	0.17	0.14	0.31	-0.48	0.47
Q1 24	0.17	0.63	0.12	-0.06	-0.09
Q1 2	0.15	-0.07	-0.27	-0.07	0.71
Q1 4	0.13	0.17	0.65	0.01	-0.15
Q1 29	0.08	0.45	0.19	0.24	0.06
Q1 21	0.07	0.73	0.04	0.05	-0.10
Q1 11	0.04	0.13	0.06	0.66	0.14
Q1 19	0.00	0.70	0.15	0.07	0.04
Q1 1	-0.02	0.37	0.41	0.14	0.16
Q1 12	-0.02	-0.17	-0.18	0.09	0.70
Q1 8	-0.18	0.13	0.19	0.23	0.59

To better visualize and interpret the factors we often “supress” loadings with small values, e.g. with absolute values smaller than 0.5. In this case our factors look as follows after suppressing the small numbers:

	Component 1	Component 2	Component 3	Component 4	Component 5
Q1 9	0.78				
Q1 26	0.72				
Q1 3	0.71				
Q1 13	0.68				
Q1 27	0.63				
Q1 28	0.62				
Q1 14	0.61				
Q1 16	0.57			0.55	
Q1 7	0.56	0.50			
Q1 17	0.55			0.54	
Q1 20	0.55				
Q1 5			0.58		
Q1 15			0.50		
Q1 23					
Q1 25					
Q1 6		0.62			
Q1 22		0.62			
Q1 18		0.73			
Q1 10					
Q1 24		0.63			
Q1 2					0.71
Q1 4			0.65		
Q1 29					
Q1 21		0.73			
Q1 11				0.66	
Q1 19		0.70			
Q1 1					

	Component 1	Component 2	Component 3	Component 4	Component 5
Q1 12					0.70
Q1 8					0.59

Questions

1. What do the first couple of factors mean? Do they make business sense?
2. How many factors should we choose for this data/customer base? Please try a few and explain your final choice based on a) statistical arguments, b) on interpretation arguments, c) on business arguments (**you need to consider all three types of arguments**)
3. How would you interpret the factors you selected?
4. What lessons about data science do you learn when doing this analysis? Please comment.
5. (Extra/Optional) Can you make this report "dynamic" using shiny and then post it on shinyapps.io? (see for example exercise set 1 and interactive exercise set 2)

Your Answers here:

Step 6: Save factor scores

We can now either replace all initial variables used in this part with the factors scores or just select one of the initial variables for each of the selected factors in order to represent that factor. Here is how to get the factor scores and how they are for the first few respondents:

	Derived Variable (Factor) 1	Derived Variable (Factor) 2	Derived Variable (Factor) 3	Derived Variable (Fac
observation 01	1.63	1.21	1.76	
observation 02	1.39	-0.09	0.04	
observation 03	1.81	-1.19	0.95	
observation 04	0.39	0.08	2.06	
observation 05	1.67	-0.20	1.49	
observation 06	0.26	0.97	0.43	
observation 07	1.06	0.97	-0.17	
observation 08	1.19	-0.05	-0.12	
observation 09	1.80	-0.76	-0.15	
observation 10	0.85	1.37	-3.42	

Questions

1. Can you describe some of the people using the new derived variables (factor scores)?
2. Which of the 29 initial variables would you select to represent each of the factors you selected?

Your Answers here:

Part 2: Cluster Analysis and Segmentation

The code used here is along the lines of the code in the session 5-6 reading ClusterAnalysisReading.Rmd. We follow the process described in the Cluster Analysis reading.

In this part we also become familiar with:

1. Some clustering Methods;
2. How these tools can be used in practice.

A key family of methods used for segmentation is what is called **clustering methods**. This is a very important problem in statistics and **machine learning**, used in all sorts of applications such as in Amazon's pioneer work on recommender systems. There are many *mathematical methods* for clustering. We will use two very standard methods, **hierarchical clustering** and **k-means**. While the "math" behind all these methods can be complex, the R functions used are relatively simple to use, as we will see.

All user inputs for this part should be selected here:

Step 1/2: Explore the data

(This was done above, so we skip it)

Step 3. Select Segmentation Variables

For simplicity will use one representative question for each of the factor we found in Part 1 (we can also use the "factor scores" for each respondent) to represent our survey respondents. These are the `segmentation_attributes_used` selected below. We can choose the question with the highest absolute factor loading for each factor. For example, when we use 5 factors with the varimax rotation we can select questions Q.1.9 (I see my boat as a status symbol), Q1.18 (Boating gives me a feeling of adventure), Q1.4 (I only consider buying a boat from a reputable brand), Q1.11 (I tend to perform minor boat repairs and maintenance on my own) and Q1.2 (When buying a boat getting the lowest price is more important than the boat brand) - try it. These are columns 10, 19, 5, 12, and 3, respectively of the data matrix `Projectdata`.

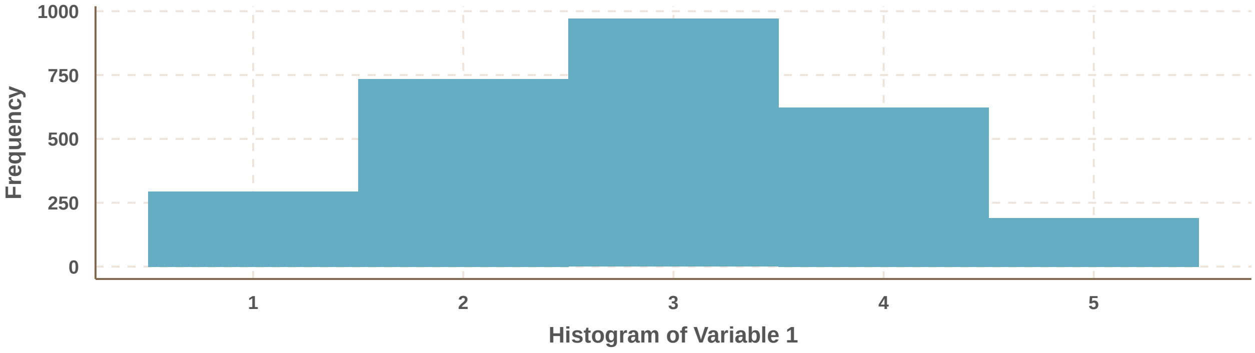
Step 4: Define similarity measure

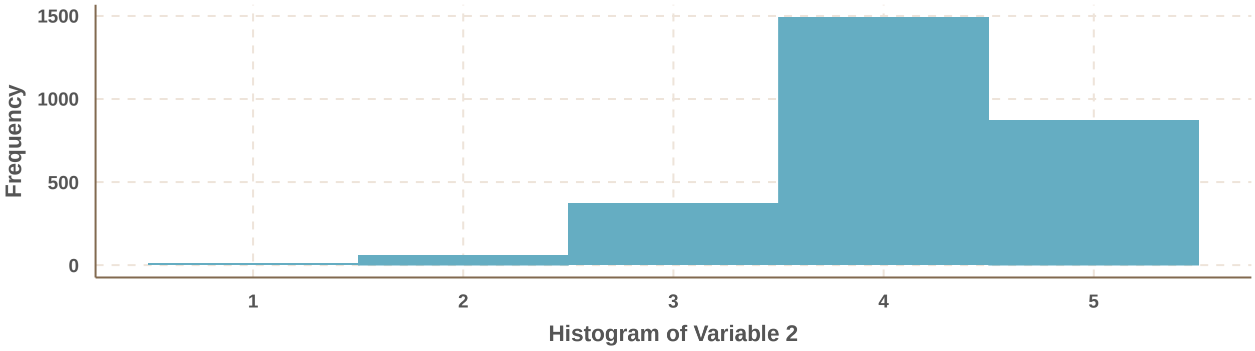
We need to define a distance metric that measures how different people (observations in general) are from each other. This can be an important choice. Here are the differences between the observations using the distance metric we selected:

1	2	3	4	5	6	7	8	9	10
0									
2	0								
2	2	0							
3	1	3	0						
3	2	4	2	0					
3	2	2	2	4	0				
4	3	4	2	3	3	0			
3	2	3	2	2	4	2	0		
4	2	4	2	2	4	1	1	0	
5	4	5	5	6	4	5	5	5	0

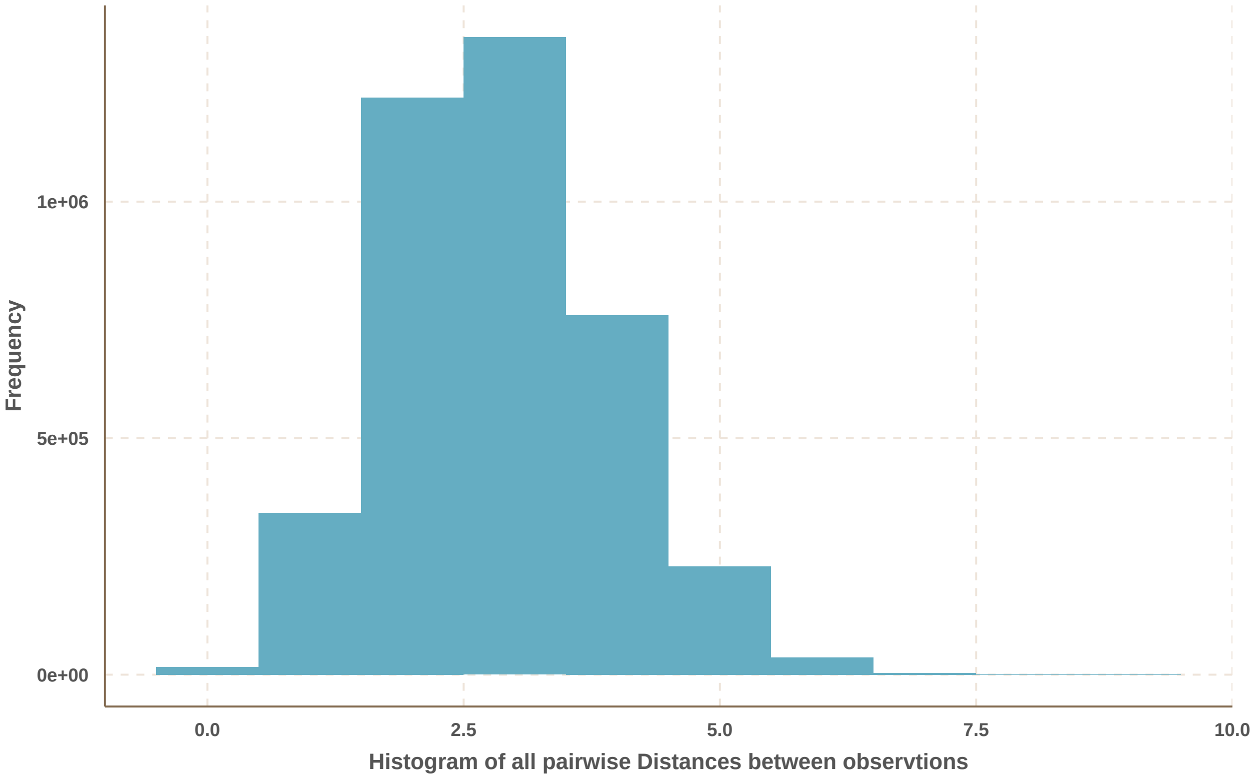
Step 5: Visualize Pair-wise Distances

We can see the histogram of, say, the first 2 variables (can you change the code below to see other variables?)



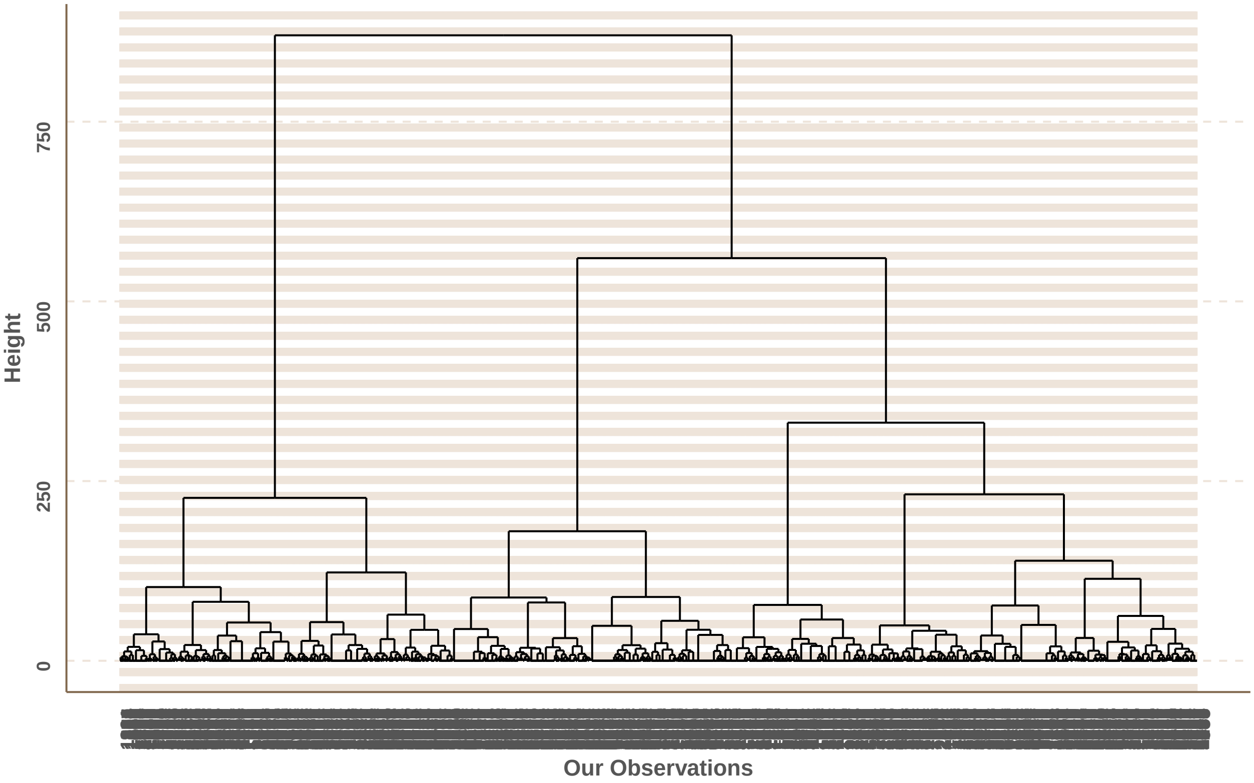


or the histogram of all pairwise distances for the euclidean distance:

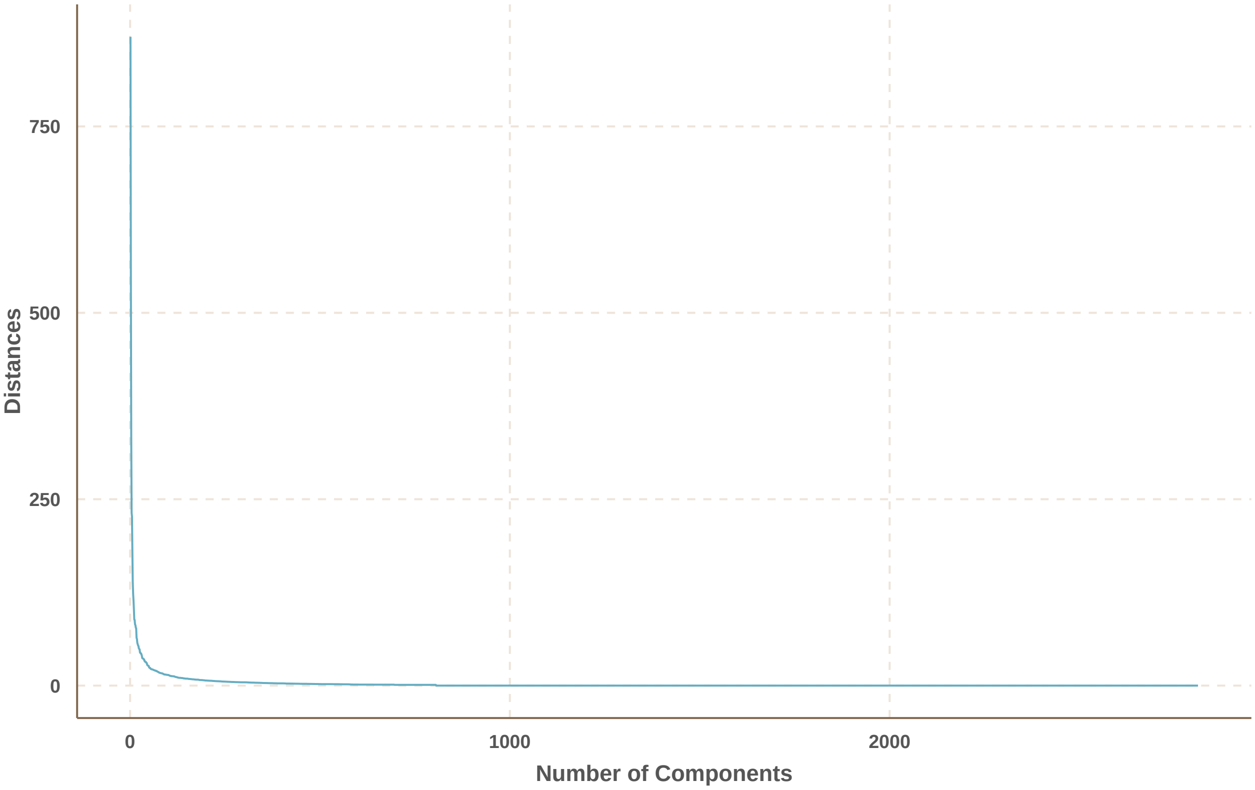


Step 6: Method and Number of Segments

We need to select the clustering method to use, as well as the number of cluster. It may be useful to see the dendrogram from Hierarchical Clustering, to have a quick idea of how the data may be segmented and how many segments there may be. Here is the dendrogram for our data:



We can also plot the "distances" traveled before we need to merge any of the lower and smaller in size clusters into larger ones - the heights of the tree branches that link the clusters as we traverse the tree from its leaves to its root. If we have n observations, this plot has n-1 numbers.



Here is the segment membership if we use hierarchical clustering:

Observation Number	Cluster_Membership
1	1
2	1
3	1
4	2
5	3
6	1
7	4
8	5
9	3
10	1

while this is the segment membership if we use k-means:

Observation Number	Cluster_Membership
1	4
2	4
3	4
4	5
5	5
6	4
7	3
8	1

Observation Number	Cluster_Membership
9	5
10	2

Step 7: Profile and interpret the segments

In market segmentation one may use variables to **profile** the segments which are not the same (necessarily) as those used to **segment** the market: the latter may be, for example, attitude/needs related (you define segments based on what the customers "need"), while the former may be any information that allows a company to identify the defined customer segments (e.g. demographics, location, etc). Of course deciding which variables to use for segmentation and which to use for profiling (and then **activation** of the segmentation for business purposes) is largely subjective. In this case we can use all survey questions for profiling for now - the `profile_attributes_used` variables selected below.

There are many ways to do the profiling of the segments. For example, here we show how the *average* answers of the respondents *in each segment* compare to the *average answer of all respondents* using the ratio of the two. The idea is that if in a segment the average response to a question is very different (e.g. away from ratio of 1) than the overall average, then that question may indicate something about the segment relative to the total population.

Here are for example the profiles of the segments using the clusters found above:

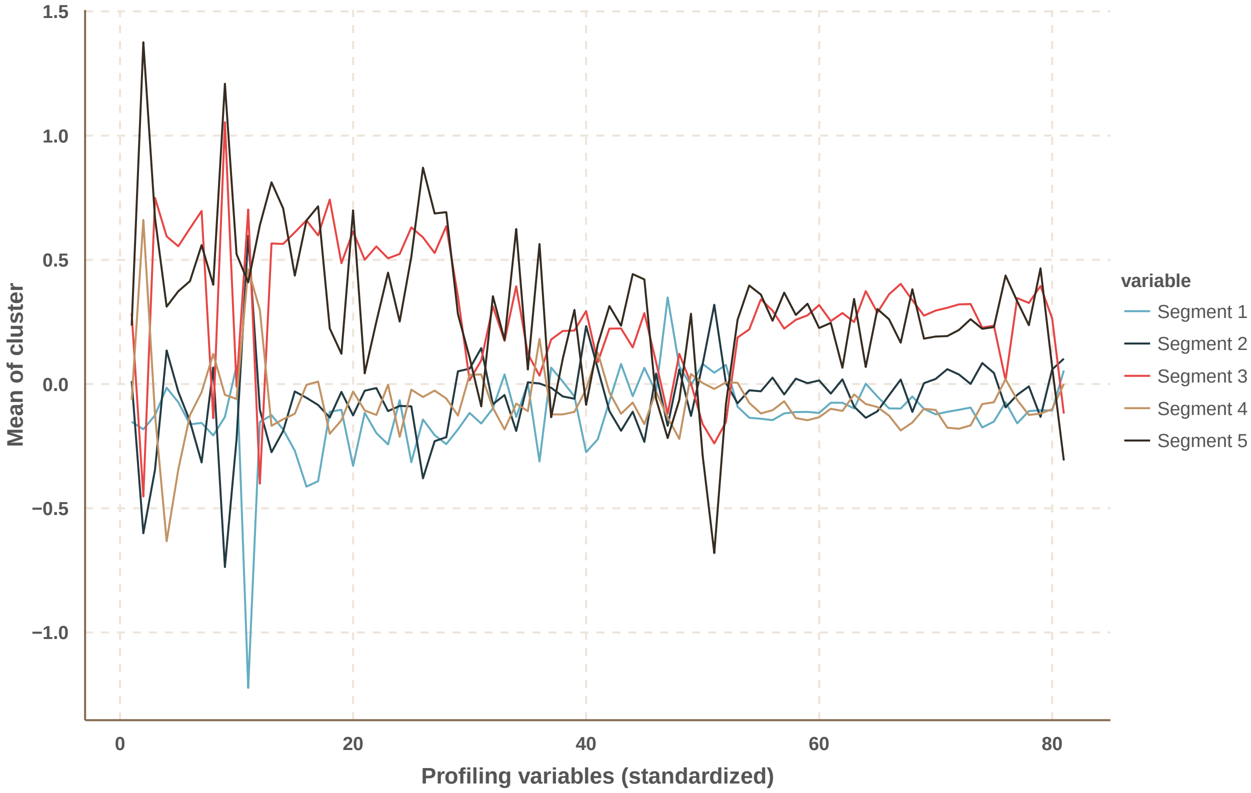
First let's see just the average answer people gave to each question for the different segments as well as the total population:

	Population	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q1 1	4.03	3.90	4.04	4.26	3.98	4.22
Q1 2	2.89	2.70	2.28	2.43	3.56	4.28
Q1 3	3.12	2.99	2.76	3.88	2.99	3.80
Q1 4	3.89	3.88	4.00	4.38	3.37	4.15
Q1 5	3.55	3.48	3.52	4.07	3.23	3.90
Q1 6	3.95	3.82	3.83	4.47	3.85	4.30
Q1 7	3.67	3.53	3.39	4.30	3.65	4.18
Q1 8	3.74	3.57	3.79	3.62	3.84	4.06
Q1 9	2.89	2.74	2.09	4.02	2.84	4.19
Q1 10	3.37	3.44	3.16	3.36	3.31	3.85
Q1 11	3.46	2.05	4.15	4.27	3.99	3.93
Q1 12	2.86	2.70	2.75	2.45	3.16	3.50
Q1 13	3.02	2.90	2.75	3.58	2.86	3.82
Q1 14	3.25	3.07	3.06	3.79	3.11	3.93
Q1 15	3.63	3.39	3.60	4.17	3.52	4.02
Q1 16	3.10	2.67	3.04	3.79	3.10	3.79
Q1 17	3.08	2.70	3.00	3.67	3.09	3.78
Q1 18	4.12	4.04	4.02	4.67	3.97	4.29
Q1 19	4.20	4.13	4.18	4.55	4.09	4.29
Q1 20	3.16	2.84	3.04	3.75	3.13	3.84
Q1 21	4.25	4.17	4.23	4.61	4.17	4.28
Q1 22	4.01	3.86	4.00	4.42	3.92	4.19
Q1 23	3.56	3.31	3.45	4.08	3.56	4.02
Q1 24	4.11	4.06	4.04	4.51	3.95	4.30
Q1 25	3.79	3.50	3.71	4.37	3.77	4.26
Q1 26	2.95	2.80	2.55	3.57	2.89	3.86
Q1 27	3.16	2.94	2.91	3.71	3.13	3.88
Q1 28	3.31	3.07	3.10	3.93	3.25	3.99

	Population	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q1 29	4.03	3.90	4.07	4.29	3.94	4.24
Q2	0.90	0.81	0.95	0.91	0.93	0.99
Q2 Cluster	0.74	0.67	0.81	0.78	0.76	0.70
Q3	4.15	4.08	4.09	4.37	4.08	4.40
Q4	3.92	4.01	3.81	4.32	3.49	4.33
Q5	3.25	3.04	2.96	3.86	3.13	4.21
Q6	22.83	22.85	22.91	24.13	21.67	23.46
Q7 1	2.23	1.89	2.24	2.27	2.44	2.86
Q7 2	4.00	4.08	3.99	4.20	3.87	3.85
Q7 3	3.80	3.81	3.75	4.04	3.66	3.92
Q7 4	3.67	3.63	3.62	3.86	3.58	3.94
Q8	2.31	2.09	2.50	2.55	2.29	2.24
Q9 1	3.57	3.30	3.65	3.68	3.73	3.77
Q9 2	3.41	3.34	3.29	3.65	3.37	3.74
Q9 3	3.72	3.80	3.54	3.93	3.60	3.94
Q9 4	3.19	3.14	3.06	3.37	3.11	3.71
Q9 5	3.51	3.57	3.26	3.81	3.34	3.95
Q10	46.25	44.80	48.12	50.19	44.95	43.64
Q11	1.45	1.62	1.37	1.39	1.38	1.34
Q12	13.42	13.76	13.70	13.99	12.40	13.13
Q13	2.08	2.07	1.85	2.08	2.15	2.59
Q14	2.27	2.42	2.42	1.96	2.28	1.72
Q15	2.54	2.60	2.91	2.27	2.52	1.76
Q16	24.77	26.34	24.85	21.66	24.90	23.12
Q16 1	3.66	3.56	3.58	3.85	3.66	3.93
Q16 2	3.56	3.44	3.54	3.75	3.49	3.90
Q16 3	3.72	3.61	3.69	3.99	3.62	4.00
Q16 4	3.76	3.64	3.78	3.99	3.67	3.96
Q16 5	3.71	3.61	3.67	3.89	3.65	4.00
Q16 6	3.82	3.73	3.84	4.03	3.71	4.04
Q16 7	3.91	3.82	3.91	4.13	3.79	4.17
Q16 8	3.91	3.82	3.92	4.15	3.80	4.08
Q16 9	3.91	3.85	3.88	4.11	3.83	4.10
Q16 10	3.83	3.77	3.85	4.07	3.74	3.89
Q16 11	3.65	3.56	3.57	3.88	3.61	3.97
Q16 12	3.56	3.56	3.44	3.90	3.49	3.62
Q16 13	3.66	3.62	3.57	3.91	3.58	3.93
Q16 14	3.75	3.67	3.71	4.05	3.65	3.97
Q16 15	3.88	3.80	3.89	4.20	3.73	4.01
Q16 16	3.67	3.62	3.57	3.97	3.53	4.01
Q16 17	3.85	3.77	3.85	4.07	3.77	4.00
Q16 18	3.88	3.79	3.90	4.11	3.80	4.03
Q16 19	3.89	3.80	3.93	4.12	3.76	4.03
Q16 20	3.97	3.90	4.00	4.20	3.84	4.13
Q16 21	3.91	3.84	3.91	4.15	3.79	4.11
Q16 22	3.93	3.80	3.99	4.09	3.87	4.09
Q16 23	3.99	3.88	4.02	4.15	3.94	4.15
Q16 24	3.31	3.24	3.22	3.33	3.33	3.72
Q16 25	3.65	3.51	3.61	3.95	3.59	3.94
Q16 26	3.90	3.81	3.89	4.14	3.80	4.07
Q16 27	3.63	3.54	3.52	3.96	3.53	4.02
Q17	0.33	0.28	0.36	0.46	0.29	0.37

	Population	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q18	0.50	0.53	0.55	0.44	0.50	0.34

We can also "visualize" the segments using **snake plots** for each cluster. For example, we can plot the means of the profiling variables for each of our clusters to better visualize differences between segments. For better visualization we plot the standardized profiling variables.



We can also compare the averages of the profiling variables of each segment relative to the average of the variables across the whole population. This can also help us better understand whether there are indeed clusters in our data (e.g. if all segments are much like the overall population, there may be no segments). For example, we can measure the ratios of the average for each cluster to the average of the population (e.g. avg(cluster)/avg(population)) and explore a matrix as the following one:

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q1 1	0.97	1.00	1.06	0.99	1.05
Q1 2	0.94	0.79	0.84	1.23	1.48
Q1 3	0.96	0.89	1.25	0.96	1.22
Q1 4	1.00	1.03	1.13	0.87	1.07
Q1 5	0.98	0.99	1.14	0.91	1.10
Q1 6	0.97	0.97	1.13	0.97	1.09
Q1 7	0.96	0.92	1.17	0.99	1.14
Q1 8	0.95	1.01	0.97	1.03	1.09
Q1 9	0.95	0.73	1.39	0.98	1.45
Q1 10	1.02	0.94	1.00	0.98	1.14

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q1 11	0.59	1.20	1.23	1.15	1.14
Q1 12	0.95	0.96	0.86	1.11	1.23
Q1 13	0.96	0.91	1.18	0.95	1.26
Q1 14	0.95	0.94	1.17	0.96	1.21
Q1 15	0.93	0.99	1.15	0.97	1.11
Q1 16	0.86	0.98	1.22	1.00	1.22
Q1 17	0.88	0.97	1.19	1.00	1.23
Q1 18	0.98	0.98	1.13	0.96	1.04
Q1 19	0.98	0.99	1.08	0.97	1.02
Q1 20	0.90	0.96	1.19	0.99	1.21
Q1 21	0.98	1.00	1.09	0.98	1.01
Q1 22	0.96	1.00	1.10	0.98	1.05
Q1 23	0.93	0.97	1.15	1.00	1.13
Q1 24	0.99	0.98	1.10	0.96	1.05
Q1 25	0.92	0.98	1.15	0.99	1.12
Q1 26	0.95	0.86	1.21	0.98	1.31
Q1 27	0.93	0.92	1.18	0.99	1.23
Q1 28	0.93	0.94	1.19	0.98	1.21
Q1 29	0.97	1.01	1.06	0.98	1.05
Q2	0.90	1.05	1.01	1.03	1.09
Q2 Cluster	0.91	1.08	1.06	1.02	0.95
Q3	0.98	0.99	1.05	0.98	1.06
Q4	1.02	0.97	1.10	0.89	1.10
Q5	0.94	0.91	1.19	0.96	1.30
Q6	1.00	1.00	1.06	0.95	1.03
Q7 1	0.84	1.00	1.02	1.09	1.28
Q7 2	1.02	1.00	1.05	0.97	0.96
Q7 3	1.00	0.99	1.06	0.96	1.03
Q7 4	0.99	0.99	1.05	0.97	1.07
Q8	0.90	1.08	1.10	0.99	0.97
Q9 1	0.92	1.02	1.03	1.04	1.06
Q9 2	0.98	0.97	1.07	0.99	1.10
Q9 3	1.02	0.95	1.06	0.97	1.06
Q9 4	0.98	0.96	1.05	0.97	1.16
Q9 5	1.02	0.93	1.09	0.95	1.13
Q10	0.97	1.04	1.09	0.97	0.94
Q11	1.12	0.94	0.96	0.95	0.93
Q12	1.02	1.02	1.04	0.92	0.98
Q13	0.99	0.89	1.00	1.03	1.24
Q14	1.07	1.07	0.86	1.00	0.76
Q15	1.02	1.14	0.89	0.99	0.69
Q16	1.06	1.00	0.87	1.01	0.93
Q16 1	0.97	0.98	1.05	1.00	1.07
Q16 2	0.97	0.99	1.05	0.98	1.10
Q16 3	0.97	0.99	1.07	0.97	1.08
Q16 4	0.97	1.01	1.06	0.98	1.05
Q16 5	0.97	0.99	1.05	0.98	1.08
Q16 6	0.98	1.00	1.05	0.97	1.06
Q16 7	0.98	1.00	1.06	0.97	1.07
Q16 8	0.98	1.00	1.06	0.97	1.04
Q16 9	0.98	0.99	1.05	0.98	1.05
Q16 10	0.98	1.00	1.06	0.98	1.01

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q16 11	0.98	0.98	1.06	0.99	1.09
Q16 12	1.00	0.97	1.09	0.98	1.02
Q16 13	0.99	0.97	1.07	0.98	1.07
Q16 14	0.98	0.99	1.08	0.97	1.06
Q16 15	0.98	1.00	1.08	0.96	1.03
Q16 16	0.99	0.97	1.08	0.96	1.09
Q16 17	0.98	1.00	1.06	0.98	1.04
Q16 18	0.98	1.00	1.06	0.98	1.04
Q16 19	0.98	1.01	1.06	0.97	1.04
Q16 20	0.98	1.01	1.06	0.97	1.04
Q16 21	0.98	1.00	1.06	0.97	1.05
Q16 22	0.97	1.02	1.04	0.99	1.04
Q16 23	0.97	1.01	1.04	0.99	1.04
Q16 24	0.98	0.97	1.00	1.01	1.12
Q16 25	0.96	0.99	1.08	0.98	1.08
Q16 26	0.98	1.00	1.06	0.98	1.05
Q16 27	0.98	0.97	1.09	0.97	1.11
Q17	0.85	1.08	1.37	0.86	1.10
Q18	1.06	1.10	0.88	1.00	0.69

Questions

1. What do the numbers in the last table indicate? What numbers are the more informative?
2. Based on the tables and snake plot above, what are some key features of each of the segments of this solution?

Your Answers here:

Step 8: Robustness Analysis

We should also consider the robustness of our analysis as we change the clustering method and parameters. Once we are confortable with the solution we can finally answer our first business questions:

Questions

1. How many segments are there in our market? How many do you select and why? Try a few and explain your final choice based on a) statistical arguments, b) on interpretation arguments, c) on business arguments (**you need to consider all three types of arguments**)
2. Can you describe the segments you found based on the profiles?
3. What if you change the number of factors and in general you *iterate the whole analysis*? **Iterations** are key in data science.
4. Can you now answer the Boats case questions? What business decisions do you recommend to this company based on your analysis?

Your Answers here:

Part 3: Classification Analysis

We will now use the classification methods to understand the key purchase drivers for boats (a similar analysis can be done for recommendation drivers). For simplicity we do not follow the "generic" steps of classification discussed in that reading, and only consider the classification and purchase drivers analysis for the segments we found above.

We are interested in understanding the purchase drivers, hence our **dependent** variable is column 82 of the Boats data (Q18) - why is that? We will use only the subquestions of **Question 16** of the case for now, and also select some of the parameters for this part of the analysis:

Questions

1. How do you select the profit/cost values for the analysis? Does the variable 100, -50, -75, 0 above relate to the final business decisions? How?
2. What does the variable 0.5 affect? Does it relate to the final business decisions? How?

Your Answers here:

We will use two classification trees and logistic regression. You can select "complexity" control for one of the classification trees here:

Question

1. How can this parameter affect the final results? What business implications can this parameter choice have?

Your Answers here:

The profit curves for the test data in this case are:

Part 4: Business Decisions

We will now get the results of the overall process (parts 1-3) and based on them make business decisions (e.g. answer the questions of the Boats case study). Specifically, we will study the purchase drivers for each segment we found.

Final Solution: Segment Specific Analysis

Let's see first how many observations we have in each segment, for the segments we selected above:

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Number of Obs.	871	739	344	592	267

This is our final segment specific analysis and solution. We can study now the purchase drivers (average answers to Q16 of the survey) for each segment. They are as follows:

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q16.2	-0.82	-0.71	-0.10	-1.00	-0.11
Q16.3	0.14	-0.25	0.27	0.03	0.00
Q16.4	-0.27	0.25	-0.37	-0.07	0.03
Q16.5	0.23	0.54	0.13	0.41	-0.30
Q16.6	0.27	-0.11	-0.23	-0.55	-0.16
Q16.7	-0.09	0.04	0.37	-0.07	0.05
Q16.8	0.05	0.04	0.27	0.48	0.35
Q16.9	0.77	0.71	0.23	-0.07	0.41
Q16.10	0.36	0.39	0.40	0.62	0.49
Q16.11	-0.73	0.14	-0.37	0.10	-0.30
Q16.12	-0.05	-0.36	0.77	0.79	1.00
Q16.13	-0.18	-0.64	-0.43	-0.07	0.19
Q16.14	-0.68	-0.18	-0.30	-0.17	0.03

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Q16.15	1.00	-0.64	0.27	0.31	0.03
Q16.16	-0.41	-0.75	-0.17	-1.00	-0.24
Q16.17	-0.59	-0.29	-0.23	-0.48	-0.19
Q16.18	-0.41	-0.43	0.10	-0.14	-0.24
Q16.19	0.77	0.32	0.13	-0.24	0.08
Q16.20	-0.05	0.00	-0.13	0.17	0.49
Q16.21	0.82	0.89	-0.30	0.72	-0.19
Q16.22	0.77	-0.25	-0.50	0.48	0.32
Q16.23	-0.73	-0.43	0.73	-0.17	0.24
Q16.24	0.23	1.00	-0.03	0.90	-0.78
Q16.25	-0.68	0.00	-0.73	0.03	0.03
Q16.26	0.64	0.86	0.73	-0.31	-0.41
Q16.27	0.68	0.43	1.00	0.00	-0.16

The profit curves for the test data in this case are:

Questions:

1. What are the main purchase drivers for the segments and solution you found?
2. How different are the purchase drivers you find when you use segmentation versus when you study all customers as "one segment"? Why?
3. Based on the overall analysis, what segmentation would you choose?
4. What is the business profit the company can achieve (as measured with the test data) based on your solution?
5. What business decisions can the company make based on this analysis?

Answers:

You have now completed your first market segmentation project. Do you have data from another survey you can use with this report now?

Extra question: explore and report a new segmentation analysis. . .