Group 3
GDMO
Assignment 2
11/11/2020

THE PROBLEM

A survey is submitted to 15 customers of a supermarket asking them if 5 different aspects of the shop should be improved, namely

- m1=quality of fresh vegetables
- m2=quality of fresh meat
- m3=quality of fresh fish
- m4=variety of products for home cleaning
- m5=variety of products for personal hygiene

The collected data are stored in the file data3.csv. Higher grades correspond to an advice of bigger improvement, while low grades mean that the customer is satisfied of the present situation.
Can you interpret the results in terms of "concepts" behind the evaluation? And can you group and classify the customers with respect to their attention to such concepts?

INDEX:

INTRODUCTION

We have a Data Matrix M (n x d) containing the grades assigned by 15 customers (n) to 5 aspects (d) of a shop which should be improved (i.e. a higher grade corresponds to higher dissatisfaction). We are asked to identify the concepts hidden in M to which the customers and the aspects of the shop to be improved are connected. We will proceed by decomposing M through the *Singular Value Decomposition* to identify the most relevant concepts; in this way we can reduce the dimension of the problem and give the best k-approximation of M (with k=number of retained concepts). We can then study the similarities among customers in order to group and classify them based on their relation to each concept.

PRELIMINARY CONSIDERATIONS

At first glance, it is hard to determine whether there is a pattern upon which customers assign grades. In order to get a better visual representation of data, we decided to divide the grades, which in our sample range from 0 (being the highest level of satisfaction) to 15 (being the highest level of dissatisfaction), into four subgroups:

- Green: grade = 0 (customers would not improve that aspect at all)
- Yellow: grades 1-5 (customer is overall satisfied)
- Orange: grades 6-10 (customer is somehow unsatisfied)
- Red: grades 11-15 (customer shows a significant level of dissatisfaction)

legend	
m1: quality of fresh vegetables	degree of dissatisfaction
m2: quality of fresh meat	0
m3: quality of fresh fish	1 to 5
m4: variety of products for home cleaning	6 to 10
m5: variety of products for personal hygiene	11 to 15

We then reordered the rows of M to try and visualize the grades' distribution (the reordering has been carried out with a custom sort in excel applied consecutively to the five columns of the matrix).

customer	m1	m2	m3	m4	m5
7	0	0	0	9	10
10	0	0	0	10	14
15	0	0	0	11	7
9	0	0	0	11	11
4	0	0	1	6	6
13	0	0	1	6	10
12	0	0	2	10	5
1	1	0	1	11	7
11	1	1	0	13	6
8	2	0	0	9	10
3	5	9	10	1	0
6	8	7	7	0	2
14	9	7	7	1	0
2	9	11	13	0	1
5	10	9	15	0	1

We notice that:

- the sample of 15 customers can be intuitively divided in two groups:
  - Customers (1, 4, 7, 8, 9, 10, 11, 12, 13, 15) who are completely or significantly satisfied with the quality of fresh vegetables, meat and fish, but not with the other categories of products
  - Customers (2, 3, 5, 6, 14) who are completely or significantly satisfied with the variety of products for home cleaning and personal hygiene, but not with the other categories of products
- m1, m2, m3 all fall in the category "alimentary products", whereas m4 and m5 fall in the category "chemical products".
- there may be strong *covariance* patterns between the first three variables (m1,m2,m3) and between the last two (m4,m5); in fact, as we will check later exploiting R, there is strong negative correlation between the two groups of products, and strong positive correlation within them.

We deduce that M connects the first three aspects (m1,m2,m3) to one concept and the other two aspects (m4, m5) to another concept, and it relates each customer to one of these concepts.
Thus, we expect the fundamental concepts behind the customers' evaluations to be two.

METHODOLOGY

MATHEMATICS OF THE PROBLEM

To get a mathematical understanding of what will follow in the next paragraph, it is necessary to point out that for the decomposition of the matrix we are using a number of left singular values, right singular values, and singular values equal to r (with r being the rank of M).
The SVD function will therefore generate:

-U: an $n \times r$ column-orthonormal matrix with each of its columns being a unit vector and the dot product of any two columns being 0, relating customers to concepts.
-Vt: an $r \times d$ column-orthonormal matrix relating aspects to concepts.
-Σ: an $r \times r$ diagonal matrix (denoted with D in this report) containing the singular values of M (σ), with all elements equal to 0 except those on the main diagonal, representing the strength of each concept.

Based on the entries of these matrices, we will try to give an interpretation of the columns of U, D, and V (the transpose of Vt). As stated in the previous paragraph, we expect the n rows and d columns of the matrix M to be connected with two main relevant concepts. Therefore, since the SVD will generate no more than five singular values, we will check whether it will be possible to omit some of the singular values while still obtaining a good approximation of M. In order to do so, we will compute the energy of D and set some of the singular values equal to 0 (hence eliminating the corresponding columns of U and V), starting from the smallest and stopping when the approximated D will have retained no less than 90% of the energy of the original one. This procedure will guarantee the rank k (with k = number of singular values retained) best approximation of M (denoted from now on with Mbest).

After reducing the dimensionality of the problem, it will be easier to extrapolate useful insights about the survey. To do so, we carry out a second SVD, setting the rank equal to the number of singular values k, and conduct an analysis of its components (Ubest, Dbest, Vbest).
First, we map the entries associated with each aspect into the matrix Vbest, with k right singular values (the transpose of such matrix is denoted from now on with Vtbest) to find the score of each customer. This can be carried out with a matrix-matrix multiplication between M and Vbest. Now we can interpret the scores, which are useful to determine the patterns of similarities among the customers in relation to the k concepts of the matrix Mbest. We then generate a similarity matrix with entries equal to the cosine of the angle between the vectors associated to each customer, which unequivocally shows the division of the customers in the two concepts-based groups.

COMPUTATION and INTERPRETATION

We proceed with the computation of the SVD in R.

#First, we open the **data** and remove the first column (customers);

```
> data<-read.csv("data3.csv")
> M<-data[-1]
```

#Then, we compute the **rank** of M and find that all columns are linearly independent, hence M is full rank with r = 5;

```
> library(Matrix)
> rankMatrix(M)
[1] 5
```

FIRST SVD

#We compute **SVD** of M according to its rank and print the matrices U, D, and Vt;

```
> svd_res<-svd(M,nu=5,nv=5)
> U<-svd_res$u
> D<-diag(svd_res$d[1:5],nrow=5,ncol=5)
> Vt<-t(svd_res$v)
```

***U** is the matrix that connects customers to concepts, displaying the weight that each customer gives to each concept while assigning grades;*
***D** is the diagonal positive matrix containing the singular values σ in its diagonal, in decreasing order from top-left to bottom-right;*
***Vt** connects concepts (rows) to variables (columns). It shows the magnitude to which each analysed variable is related to each concept.*

```
> U
            [,1]        [,2]        [,3]         [,4]         [,5]
 [1,] -0.3019426 -0.07212268 -0.26542535 -0.029947880 -0.197483432
 [2,] -0.1456410  0.51369493  0.04248755  0.143258585  0.257216014
 [3,] -0.1116086  0.37438055 -0.12181804  0.416575757  0.528588733
 [4,] -0.1985573 -0.04954223  0.04576741  0.135696508 -0.089763799
 [5,] -0.1519369  0.53430957  0.05135900  0.223090328 -0.663794970
 [6,] -0.1163549  0.32805616  0.15336405 -0.425231350  0.122872418
 [7,] -0.3063881 -0.10647357  0.15844210  0.025034167  0.085109978
 [8,] -0.3138158 -0.07916004  0.16263024 -0.297905728 -0.120101655
 [9,] -0.3555478 -0.12398650  0.08598286  0.025745964  0.089965538
[10,] -0.3851665 -0.13283529  0.43053619  0.039282573  0.127794095
[11,] -0.3188408 -0.08305440 -0.51745826 -0.157825505  0.193571090
[12,] -0.2550431 -0.05139110 -0.36456087  0.246865493 -0.253394022
[13,] -0.2604986 -0.06966252  0.39813736  0.151573660 -0.043756557
[14,] -0.1059354  0.34553164 -0.10100271 -0.596268619 -0.006060145
[15,] -0.2936066 -0.10386621 -0.26638709  0.009868812  0.043958296
```

```
> D
          [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 41.90228  0.00000 0.000000 0.000000 0.000000
[2,]  0.00000 35.54952 0.000000 0.000000 0.000000
[3,]  0.00000  0.00000 8.382784  0.000000  0.000000
[4,]  0.00000  0.00000 0.000000  4.945197  0.000000
[5,]  0.00000  0.00000 0.000000  0.000000  3.115207
```

```
> Vt       [,1]         [,2]         [,3]         [,4]        [,5]
[1,] -0.15562053 -0.139582724 -0.193678977 -0.705517899 -0.64886968
[2,]  0.48549147  0.519301370  0.642975961 -0.221879756 -0.17881669
[3,]  0.01755415 -0.037898770 -0.009492094 -0.672935231  0.73846030
[4,] -0.79850076  0.004957114  0.601599360 -0.008054469  0.01962891
[5,] -0.31963834  0.842248099 -0.432502576 -0.010352223  0.03583052
```

SINGULAR VALUES SELECTION

#We compute the **correlation matrix** to look for the previously supposed correlation pattern.

```
> library(corrplot)
corrplot 0.84 loaded
> Mcor<-cor(M)
> corrplot(Mcor,method="number")
```

Given the result, we can notice strong positive correlation within the set of aspects m1,m2,m3 and within the set m4 and m5, and strong negative correlation between m1,m2,,m3 and m4,m5. This is an indicator of the fact that the first three variables convey the same type of information, and the same holds for the last two.

#We create a 1×5 vector containing the singular values on the diagonal of D.

```
> singluar_values<-svd_res$d
> singular_values
[1] 41.902278   35.549522   8.382784   4.945197   3.115207
```

From D we notice that the first two singular values are very high compared to the others, thus they contain the majority of the variability of data. This is due to the fact that the higher is σ, the closer is the data to the relative axis in the n dimensional space, the more the variability along that dimension.

The most useful concepts are those displaying the biggest variability within data, as they help distinguish and explain the customers' behaviour.

The difference in the values of the five diagonal entries of D is due to the differences in the number of customers and the grades assigned to the aspects captured by the each singular value.

In particular we notice that concept 1 and 2 display values significantly higher than the others (41.90228 and 35.54952).

To confirm our feeling we check the percentage of energy retained in the system by those two singular values alone; the retained energy when excluding concepts should always be at least the 90% of the total energy, which is computed through squared values.

#We compute the sum of the square of the singular values to calculate the total **energy** of D, then proceed by eliminating the singular values one by one starting from the smallest and iterating the operation.

```
> squared_values<-singular_values^2
> D_Energy_tot<-sum(squared_values)
> D_Energy_tot
[1] 3124
> energy_retained<-sum(squared_values[1:2])/D_Energy_tot
> energy_retained
[1] 0.9665715
```

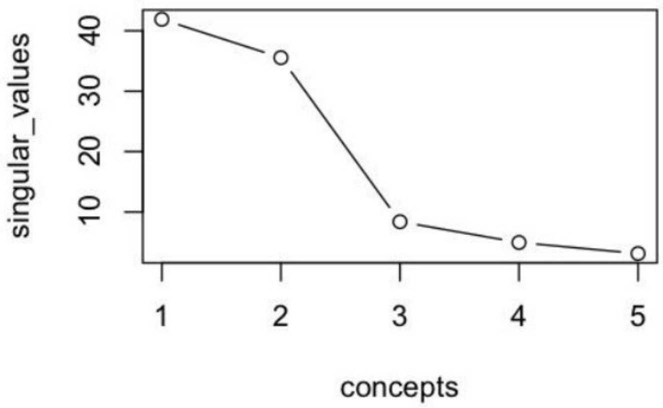
We find that we can retain up to around 96.7% of the energy of D by keeping only k=2 singular values, while for k>2 the energy of the approximated D drops drastically below 90%. For the sake of simplicity, the previous code only shows the passage where we retain two singular values.

We obtain the same result by plotting the singular values against the number of concepts (the scree plot generated allows a better visual understanding of the composition of the energy of D).

We use the "elbow rule" to find the smallest number of concepts to be retained: we see that after the second concept, the marginal variance explained by the introduction of a new one drastically reduces, and then gets flat.

#We plot the singular values against the 5 concepts related to them.

```
> concepts<-c(1,2,3,4,5)
> singular_values<-svd_res$d
> plot(concepts,singular_values,type="b")
```



SECOND SVD

Based on the result of the previous passage, we set the last 3 singular values to 0 and eliminate the relative columns and rows of U and Vt. We recompute the SVD with k = 2 to obtain the best approximation of M (Mbest). The components of the decomposition Ubest, Dbest and Vtbest are now easier to interpret since they only provide us with the most relevant information for our purpose.

#We proceed with the second **SVD** considering rank=2 and print the resulting matrices.
We specify the dimensions of Dbest since the software automatically computes it with dimensions equal to the minimum dimension of the starting matrix, which is the number of columns=5, and highlight the larger absolute values for each row of Ubest and each column of Vtbest;

```
> svdBest<-svd(M,nu=2,nv=2)
> Ubest<-svdBest$u
> Dbest<-diag(svdBest$d[1:2],nrow=2,ncol=2)
> Vtbest<-t(svdBest$v)
> Ubest
```

	[,1]	[,2]
[1,]	**-0.3019426**	-0.07212268
[2,]	-0.1456410	**0.51369493**
[3,]	-0.1116086	**0.37438055**
[4,]	**-0.1985573**	-0.04954223
[5,]	-0.1519369	**0.53430957**
[6,]	-0.1163549	**0.32805616**
[7,]	**-0.3063881**	-0.10647357
[8,]	**-0.3138158**	-0.07916004
[9,]	**-0.3555478**	-0.12398650
[10,]	**-0.3851665**	-0.13283529
[11,]	**-0.3188408**	-0.08305440
[12,]	**-0.2550431**	-0.05139110
[13,]	**-0.2604986**	-0.06966252
[14,]	-0.1059354	**0.34553164**
[15,]	**-0.2936066**	-0.10386621

From Ubest we confirm the assumption we made about the customers' division into two groups based on the strongest connection to a concept rather than to the other. We observe that customers {1,4,7,8,9,10,11,12,13,15} all show a higher absolute value in relation to the first concept (column), which means they are mainly connected to it. The same holds for customers {2,3,5,6,14}, which show a higher absolute value in relation to the second concept.

```
> Dbest
         [,1]     [,2]
[1,] 41.90228  0.00000
[2,]  0.00000 35.54952

> rownames(Vtbest)<-c("chemical","alimentary")
> colnames(Vtbest)<-c("m1","m2","m3","m4","m5")
> Vtbest
                   m1         m2        m3         m4         m5
chemical   -0.1556205 -0.1395827 -0.193679 -0.7055179 -0.6488697
alimentary  0.4854915  0.5193014  0.642976 -0.2218798 -0.1788167
```

The elements of Vtbest are a measure of the connection between each aspect that needs improvement and the two concepts respectively. The first row displays higher absolute values in the last two columns (corresponding to m4,m5), meaning that it is more related with personal hygiene and home cleaning products. Reversely, the second row shows higher absolute values in relation to the products fresh vegetables, fresh meat and fresh fish, corresponding to the first three columns(m1,m2,m3). Thus, we can associate Concept 1 to the category "chemical products" and Concept 2 to "alimentary products"

#We compute Mbest, print it, and check its rank;

```
> Mbest<-Ubest%*%Dbest%*%Vtbest
> Mbest
            [,1]       [,2]         [,3]       [,4]       [,5]
 [1,]  0.7241595  0.4345617  0.801899493  9.4951553  8.6680267
 [2,]  9.8155595 10.3351089 12.923738630  0.2536764  0.6943708
 [3,]  7.1892135  7.5641879  9.463168634  0.3464550  0.6546594
 [4,]  0.4397146  0.2467339  0.478998960  6.2606873  5.7135311
 [5,] 10.2124021 10.7524978 13.446032590  0.2771987  0.7345069
 [6,]  6.4206513  6.7367576  8.442828506  0.8521627  1.0781841
 [7,]  0.1602857 -0.1735867  0.052801577  9.8975250  9.0072575
 [8,]  0.6801251  0.3740900  0.737401115  9.9016696  9.0355844
 [9,]  0.1785939 -0.2093641  0.051460434 11.4889622 10.4551947
[10,]  0.2190119 -0.1994860  0.089572555 12.4343710 11.3167515
[11,]  0.6456792  0.3315865  0.689166083 10.0809376  9.1969630
[12,]  0.7761403  0.5429778  0.895153549  7.9451475  7.2610813
[13,]  0.4963685  0.2375789  0.521789302  8.2505472  7.5255601
[14,]  6.6543164  6.9984291  8.757713739  0.4062912  0.6837980
[15,]  0.1219400 -0.2002091  0.008670092  9.4991023  8.6431656
> rankMatrix(Mbest)
[1] 2
```

We multiplied Ubest, Dbest and Vtbest to find the best approximation matrix of M (Mbest). By finding the rank of Mbest we confirmed that the data of the original matrix M conveys most information along 2 of its 5 dimensions. As seen in the decomposition, Mbest relates customers to the different types of products according to their level of dis/satisfaction towards two concepts, (1)"chemical" and (2)"alimentary" products. These are the concepts that retain most information, thus the reduction of the dimensionality of the problem does indeed yield the best approximation of the original matrix M.

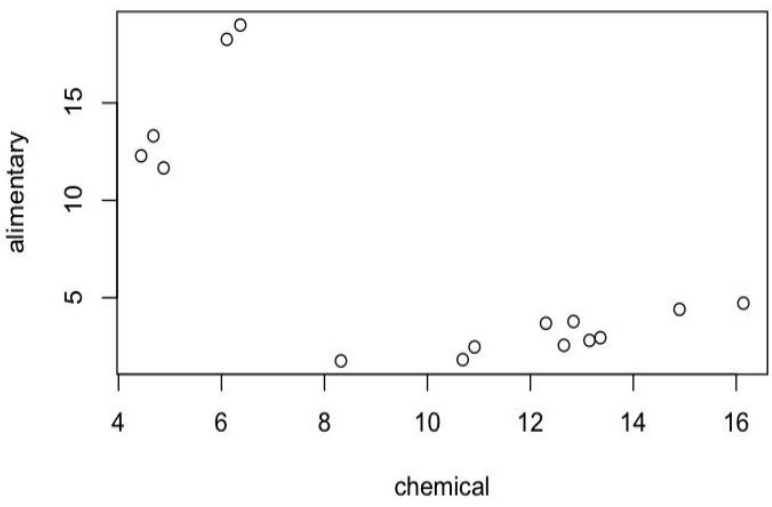
SIMILARITY ANALYSIS

To ultimately group and classify the customers we carry out a vector analysis to locate them in the "concept space" Vbest, computing their scores and analyzing the alignment of each customer's behaviour to the others.
We can compute the matrix of the scores of each customer relatively to the two concepts by multiplying the Data Matrix M by the "concept space" Vbest (note that we use the transpose of Vtbest so that dimensionalities agree).
The scores can then be used to determine the patterns of similarity among the customers in relation to the k concepts of the matrix Mbest. To do so we use the dot product to generate a square similarity matrix (with number of rows and columns equal to the number of customers=15) with entries equal to the cosine of the angle between the vectors associated to each customer.

#First, we need to express the list M in a matrix form, so that we can compute the matrix-matrix multiplication. Also, we take the absolute values of the scores and plot them to better visualize whether each customer is more unsatisfied with alimentary or with chemical products;

```
> Mm<-as.matrix(M,nrow=15,ncol=5)
> Scores<-Mm%*%Vbest
> colnames(Scores)<-c("chemical","alimentary")
> Scores
```

	chemical	alimentary
[1,]	**-12.652084**	-2.563927
[2,]	-6.102691	**18.261609**
[3,]	-4.676655	**13.309050**
[4,]	**-8.320004**	-1.761203
[5,]	-6.366504	**18.994450**
[6,]	-4.875535	**11.662240**
[7,]	**-12.838358**	-3.785085
[8,]	**-13.149599**	-2.814102
[9,]	**-14.898263**	-4.407661
[10,]	**-16.139355**	-4.722231
[11,]	**-13.360154**	-2.952544
[12,]	**-10.686885**	-1.826929
[13,]	**-10.915483**	-2.476469
[14,]	-4.438935	**12.283485**
[15,]	**-12.302785**	-3.692394




```
> S_abs<-abs(Scores)
> plot(S_abs)
```

In the score matrix 15 x 2, we write in bold the number of each row with highest absolute value. For the sake of clarity, we recall that in the original matrix M, the higher the grade assigned by a customer, the more unsatisfied the customer is with that particular product. From the score matrix we can therefore identify two groups of customers:

- *Group 1: customers 1, 4, 7, 8, 9, 10, 11, 12, 13, 15 have higher degrees of dissatisfaction with the variety of chemical products.*
- *Group 2: customers 2, 3, 5, 6, 14 have higher degrees of dissatisfaction with the quality of alimentary products*

- Notice that no 0 appears in the matrix, meaning that overall, no customer is completely satisfied with either class of products.

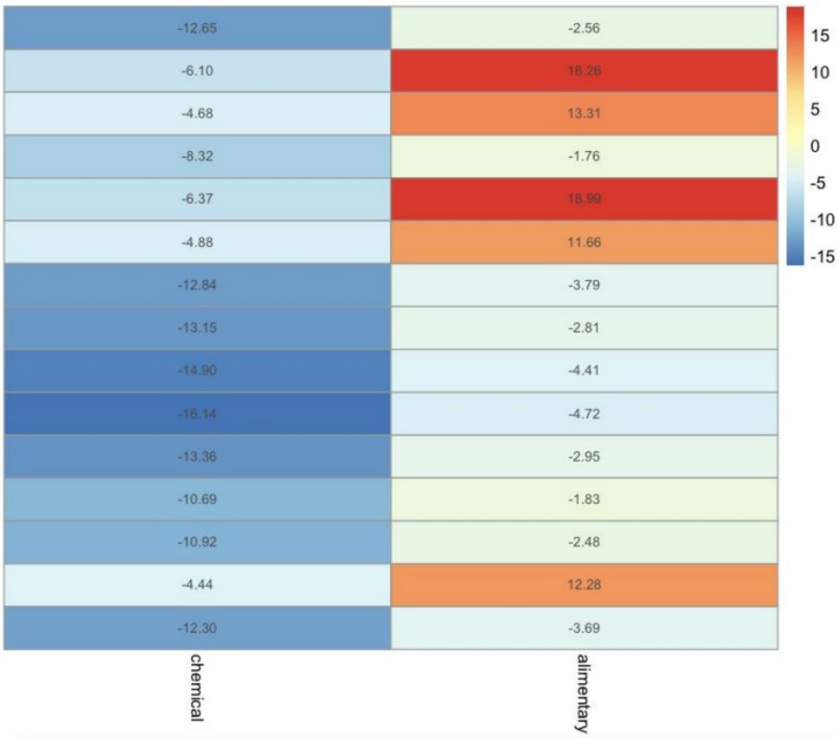
- The highest level of dissatisfaction is associated with customer 2 and 5, in relation to the alimentary products.

- The lowest level of dissatisfaction is associated with customer 4, in relation to chemical products.

- In general, group 1 is the one with highest scores, indicating that over the sample of 15 people, the majority of them are not satisfied with the variety of chemical products.

#The following visual representation of the scores is also a useful tool for interpreting results.

>pheatmap(Scores,how_colnames=TRUE,show_colnames=TRUE,cluster_rows=FALSE, cluster_cols=FALSE,display_numbers=TRUE,legend=TRUE)

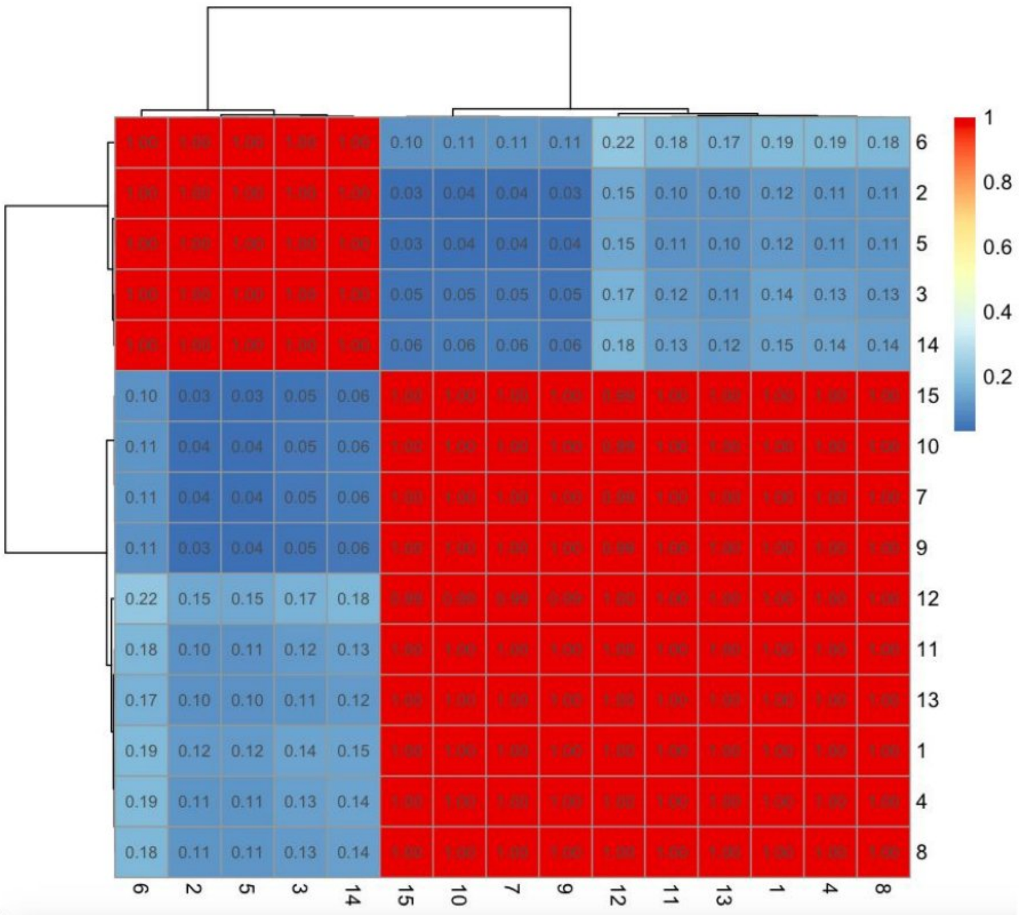


We see the scores of each customer colored with different shades and deduce the following: for each column, different shades represent the intensity according to which the customers are dis/satisfied with each concept.

Light colors express satisfaction, darker ones dissatisfaction.

#We now compute a similarity matrix on the basis of what is explained in "Methodology". Notice that R approximates to two digits.

```
> library(pracma)
> Similarity<-matrix(,nrow=15,ncol=15)
> S<-Scores
> for(i in c(1:15)){for(j in c(1:15)){Similarity[i,j]<-dot(S[i,],S[j,])/(sqrt(sum(S[i,]^2)*sum(S[j,]^2)))}}
> pheatmap(Similarity,how_colnames=TRUE,show_colnames=TRUE,cluster_rows=TRUE,
cluster_cols=TRUE,display_numbers=TRUE,legend=TRUE,labels_row=c(1:15),labels_col=c(1:15))
```



Notice that:

-The similarity index ranges from 0 to 1

-The closest the value of an entry to 0, the less similarity there is between two customers

-The closest the value of an entry to 1, the more the similarity between two customers

Results:

-The two groups identified before are meaningful, as the customers within them are almost completely similar to one another (the entries close or equal to 1)

-Among the 15 customers, Customer 15, 10, 7, and 9 (all belonging to group 1) are those who exhibit the most dissimilar behavior comparing to the people belonging to the other group

-Customer 12 (group 1) is the only one that slightly diverges from the behavior of his/her group,as she displays entries equal to 0.99 in relation with members 9, 7, 10, and 15.

CONCLUSIONS

We managed to reduce the dimension of our matrix from 5 to 2. We were able to find the best approximation of M by eliminating the "concepts" that played the least important role in how customers assigned the grades. We were left with the two most significant ones,

- *Concept 1* = "Chemical Products"
- *Concept 2* = "Alimentary Products"

and we analyzed the behaviour of our customers in relation to those.

In this way we managed to cluster the customers into two groups, based on the similarities among them. Namely

- *Group 1*, containing those customers who were unsatisfied with the variety of chemical products and were overall satisfied with the quality of alimentary products. (made up by customers *1, 4, 7, 8, 9, 10, 11, 12, 13, 15*)
- *Group 2*, containing those customers who were unsatisfied with the quality of alimentary products and were overall satisfied with the quality of chemical products (made up by customers *2, 3, 5, 6, 14)*

Our results suggest that, overall, the primary concern of the store should be to improve those aspects related to the variety of Chemical products, to which SVD assigns the biggest singular value. In fact, from the analysis of the scores we found that there is a bigger number of customers with high scores associated with Concept 1. This means that the intensity of their dissatisfaction with the variety of Chemical products is higher compared to that of members of Group 2 towards the quality of Alimentary products.
It is worth noticing that the highest degree of dissatisfaction is expressed towards the quality of Alimentary products by members 2 and 5 of Group 2. However, Group 1, which includes 10 out of the 15 customers, shows on average the highest level of dissatisfaction.
Based on our results, the aspects of the store that need the biggest improvement are those associated with the variety of Chemical products, namely personal hygiene and home cleaning products. To conclude, it is worth mentioning that our results are based on a relatively small number of customers. Therefore, they are not necessarily relevant to deduce the average attitude of the entire population of customers of this particular shop.