Assignment 2 Report

Group 5

November 8, 2021

A survey is submitted to 15 customers of a DIY (do it yourself) center asking them if 5 different aspects of the center should be improved, namely

- m1 = variety of taps and hydraulic equipments
- m2 = variety of paints
- m3 = variety of home furnishings
- m4 = speed of service at the desk
- m5 = availability of installation services

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?

Below is the original matrix representing the customers' answers to each aspect of the survey. The first column contains the answers to the first question (m1), the second column answers the second question (m2), and so on until the fifth column, which contains the answers to the fifth question. (m5).

Note that before even looking at the results, just by how the questions are stated, they can be easily divided into two groups: the first ones (m1 to m3) are questions regarding the variety of products, and the others (m4 and m5) regard customers service.

$$\begin{array}{c} \begin{array}{ccccc} m1 & m2 & m3 & m4 & m5 \end{array} \\ \left(\begin{array}{ccccc} 1 & 0 & 1 & 12 & 8 \\ 2 & 1 & 0 & 12 & 11 \\ 0 & 1 & 2 & 5 & 11 \\ 13 & 10 & 10 & 1 & 3 \\ 1 & 0 & 0 & 11 & 11 \\ 0 & 0 & 2 & 15 & 10 \\ 8 & 15 & 3 & 2 & 1 \\ 1 & 0 & 1 & 11 & 9 \\ 12 & 8 & 13 & 1 & 0 \\ 0 & 0 & 1 & 6 & 16 \\ 9 & 11 & 12 & 1 & 1 \\ 1 & 1 & 0 & 10 & 10 \\ 1 & 1 & 0 & 8 & 8 \\ 0 & 0 & 1 & 10 & 9 \\ 19 & 7 & 11 & 0 & 0 \end{array} \right) \end{array}$$

Figure 1: 15×5 matrix - survey answers.

Now, we proceed to check the rank of the matrix, to see if there's redundant information in it:

$$\text{Rank(M)} = 5$$

Since the rank of the matrix can be interpreted as the number of linearly independent vectors of the matrix, we can establish there is no linear dependence between the consumers' opinions and thus there also is no redundant information in the matrix.

To provide a better and simplified understanding of the customers' opinions, we run a singular value decomposition (SVD). The result of the SVD computation are three matrices, each representing a specific relation.

The first matrix of the SVD, here called U matrix, relates each customer to the concepts contained in the dataset. The customers are represented by the rows of the matrix.

$$\begin{pmatrix} 0,285 & 0,117 & 0,075 & 0,3 & 0,017 \\ 0,331 & 0,124 & 0,113 & 0,05 & 0,281 \\ 0,235 & 0,075 & 0,212 & 0,38 & 0,276 \\ 0,213 & 0,433 & 0,07 & 0,108 & 0,52 \\ 0,308 & 0,145 & 0,021 & 0,009 & 0,18 \\ 0,354 & 0,154 & 0,055 & 0,384 & 0,317 \\ 0,163 & 0,343 & 0,805 & 0,302 & 0,072 \\ 0,285 & 0,117 & 0,016 & 0,168 & 0,038 \\ 0,171 & 0,452 & 0,296 & 0,228 & 0,097 \\ 0,308 & 0,145 & 0,327 & 0,637 & 0,031 \\ 0,18 & 0,431 & 0,047 & 0,008 & 0,617 \\ 0,285 & 0,117 & 0,085 & 0,025 & 0,097 \\ 0,23 & 0,088 & 0,08 & 0,025 & 0,1 \\ 0,266 & 0,124 & 0,008 & 0,093 & 0,147 \\ 0,134 & 0,391 & 0,261 & 0,127 & 0,043 \end{pmatrix}$$

Figure 2: U matrix - SVD. The entries are considered in absolute value.

The second matrix of the SVD, here called the D matrix, is a diagonal matrix representing the strength of the concepts that emerge from the analysis. The non-zero entries of this matrix are the singular values of the original matrix.

$$\begin{pmatrix} 47.249 & & & & \\ & 38.065 & & & \\ & & 10.035 & & \\ & & & 9.663 & \\ & & & & 4.042 \end{pmatrix}$$

Figure 3: D matrix - SVD.

The third matrix of the SVD, called the V matrix, relates the concepts emerging from the analysis, on the columns, to the objects of the original matrix (in this case, the five questions asked in the survey). Anyway, in most calculations needed for SVD analysis, V is considered in its transposed version, usually denoted as V^T, in which concepts are found on the rows and the objects of the original dataset on the columns.

$$\begin{pmatrix} 0,236 & 0,545 & 0,057 & 0,08 & 0,798 \\ 0,211 & 0,53 & 0,669 & 0,33 & 0,343 \\ 0,229 & 0,519 & 0,612 & 0,248 & 0,491 \\ 0,645 & 0,274 & 0,306 & 0,642 & 0,047 \\ 0,656 & 0,278 & 0,282 & 0,641 & 0,04 \end{pmatrix}$$

Figure 4: V matrix - SVD. Entries are in abolute value.

In order to approximate the data in an effective way, we can start by computing the original and the partial energy of the D matrix: as long as the partial energy is at least 90 % of the original one, we can set the remaining singular values to 0, thus reducing the dimensions of the SVD.

Original energy $= 47.249^2 + 38.065^2 + 10.035^2 + 9.663^2 + 4.042^2 = 3782.268$

Partial energy (2 main values) $= 47.249^2 + 38.065^2 = 3681.412$

Partial energy/Original energy $= \frac{3681.412}{3892.999} = 0.943$

As shown by the computations, the partial energy of the two biggest singular values represents 94,3% of the original energy, meaning that the three smallest singular values can be dropped. Thus, the best possible approximation of the original matrix can be made using merely two dimensions.

Now that we know we can work with a rank 2 matrix, we proceed to compute the SVD of this more efficient matrix. We will call these three new matrices U_2, D_2 and V_2, respectively.

Using the U_2 matrix, we can now easily identify that the customers who answered the survey can be classified into two groups according to their qualifications of the concepts. The first column of this matrix represents the answers regarding the concept "customer service", and the second column represents the concept "variety of goods".

Recall that the larger the absolute value, the most dissatisfied is the customer with the concept. Based on this information, we can assume that the strongest concept is actually the "customer service" one; this assertion is also corroborated in the ${V_2}^{\mathrm{T}}$ matrix, since in the first row we have higher absolute values in the last two columns, which are the "customer service" columns. Thus, we can assume the same concept is expressed in the first column of the U_2 matrix, and that it also correspond to the first singular value of the D_2.

$$\begin{array}{cc} Service & Variety \\ \left(\begin{array}{cc} 0,284895815 & 0,116864901 \\ 0,331181119 & 0,124134413 \\ 0,235260786 & 0,075071722 \\ 0,213364708 & 0,43273625 \\ 0,308065373 & 0,145174869 \\ 0,353506326 & 0,15376512 \\ 0,162603224 & 0,342608533 \\ 0,285128336 & 0,116951537 \\ 0,172316041 & 0,453360136 \\ 0,309076013 & 0,146300707 \\ 0,179736203 & 0,431198235 \\ 0,284971198 & 0,116756204 \\ 0,22986821 & 0,0877555 \\ 0,266469908 & 0,124073681 \\ 0,134507398 & 0,39071799 \end{array}\right) \end{array}$$

Figure 5: U_2 matrix - SVD. Entries are in absolute values.

The D_2 matrix is shown below. As expected, it still holds the 2 larger singular values of the rank 5 D matrix.

$$\begin{array}{c} \begin{array}{cc} \mathit{Service} & \mathit{Variety} \end{array} \\ \begin{pmatrix} 47.249 & \\ & 38.065 \end{pmatrix} \end{array}$$

Figure 6: D_2 matrix - SVD.

The matrix $V_2{}^T$ transposed relates the concepts of the survey to the questions asked. The concepts are on the rows, and the questions are on the columns. As stated earlier, the first row is linked to the last two questions ("consumer service", in short "services"), while the second row is linked to the first three questions ("variety of goods", in short "variety").

$$\begin{array}{c} \\ Services \\ Variety \end{array} \begin{array}{c} \begin{array}{ccccc} m1 & m2 & m3 & m4 & m5 \end{array} \\ \begin{pmatrix} 0,236198933 & 0,210607384 & 0,229018688 & 0,645407899 & 0,656394453 \\ 0,545446956 & 0,529813742 & 0,51908031 & 0,274335848 & 0,277633735 \end{pmatrix} \end{array}$$

Figure 7: $V_2{}^T$ matrix - SVD. Entries in absolute value.

We can also compute the Score Users matrix, which maps each customer in a two-concept (two dimensional) space. This matrix is useful when it comes to classifying individuals with regards to their opinions on the areas of improvement so as to have some form of profiling for the business.

$$\begin{array}{cc} Services & Variety \\ \left(\begin{array}{cc} 13,46126803 & 4,448572789 \\ 15,64823902 & 4,725293608 \\ 11,11602324 & 2,857675965 \\ 10,08143811 & 16,4725139 \\ 14,55602481 & 5,526218458 \\ 16,70310039 & 5,853214449 \\ 7,682968541 & 13,04171728 \\ 13,47225459 & 4,451870677 \\ 8,141897114 & 17,2575816 \\ 14,60377733 & 5,569074541 \\ 8,492498233 & 16,41396791 \\ 13,46482984 & 4,444435133 \\ 10,86122513 & 3,340495967 \\ 12,59064776 & 4,722981785 \\ 6,355446589 & 14,87304917 \end{array}\right) \end{array}$$

Figure 8: Score Users matrix. Entries are in absolute value.

In order to provide a more intuitive understanding of the similarity between the users, we use the first customer's answers as base and compare it to the other fourteen users.

This gives us a similarity vector where each row represent one of the customers: the correspondent value, defined as the similarity coefficient, is the degree to which each customer is similar to the first one. This coefficient varies between 0 and 1 because it is the cosine of theta, where theta is the angle between the vectors of the users as defined in the score users matrix (fig.8).
The higher the value of the similarity coefficient, the higher the similarity between the answers of the customers. The maximum possible value, and thus degree of similarity, is 1, which means there is perfect equality, as it happens when comparing the first customer with himself, or nearly perfect equality, as it happens when comparing the first customer' answers with the ones provided by the eighth customer.

$$\begin{pmatrix} 1 \\ 0,999664346 \\ 0,997719716 \\ 0,228011226 \\ 0,999046645 \\ 0,999840124 \\ 0,211586903 \\ 1 \\ 0,12135028 \\ 0,998980984 \\ 0,157632838 \\ 0,999999937 \\ 0,99978382 \\ 0,999212013 \\ 0,084553843 \end{pmatrix}$$

Figure 9: Similarity vector.

Finally, we use a scatterplot to provide an immediate graphical representation, based on the Score Users matrix, of the classification in two groups of the customers with regards to their opinions. The two groups show how one subset of customers believe it is more urgent to improve on variety of goods while the other one believe it is more urgent to improve on customer service.

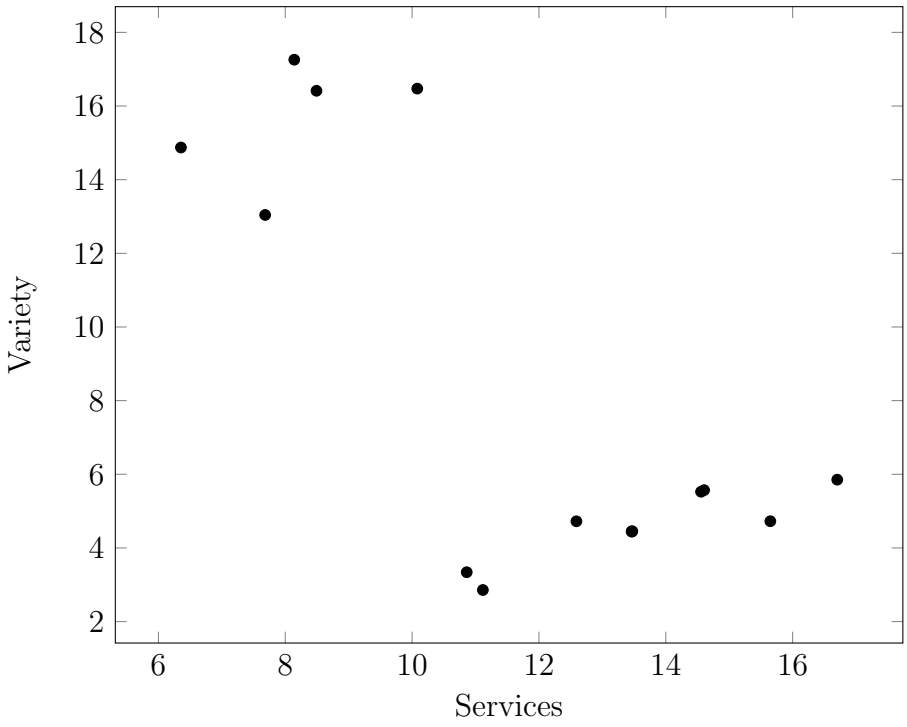

Figure 10: Scatterplot of the Score Users matrix