SemeAd 2023
XXVI Seminários em Administração

07, 08, 09 e 10 de novembro de 2023
ISSN 2177-3866

Application of unsupervised machine learning techniques in the development of intralogistics automation projects

DENIS TOYOSHIMA
ESCOLA DE ENGENHARIA DE SÃO CARLOS - EESC

FERNANDO FREIRE VASCONCELOS
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

Application of unsupervised machine learning techniques in the development of intralogistics automation projects

Introduction

Warehouses, also known as distribution centers, play a key role in the supply chain. Its main functions are the balancing of the flow of materials between production and demand, the consolidation of orders for optimization of transportation and value-added activities such as kit formation, labeling and customization (Gu et al., 2007).

Intralogistics represents the organization and optimization of logistics technologies and internal material flow of the warehouse (Zrnic et al., 2021). It plays an important role in the global industry through the automation of operations and process optimization (Fernandes et al., 2019).

The development of intralogistics projects, also known as warehouse projects, involves five main aspects: determination of the overall structure of the warehouse, sizing of the warehouse and its departments, elaboration of a detailed layout for each department, selection of equipment and selection of operational strategies. These aspects are strongly correlated and directly affect the operational efficiency of the warehouse (Gu et al., 2010).

The main processes in a warehouse are: receiving, storing, picking orders and shipping (Rouwenhorst et al., 2000). Among these processes, the separation of orders accounts for more than half of the operating costs and, in the execution of this activity, most of the time is consumed with the displacement to the address of the product (Frazelle, 2002). There are four approaches to reducing travel time in order picking: optimizing order routing; warehouse zoning; separation of orders by lots; optimization of the allocation of products in the storage positions. The fourth approach has the greatest influence on improving the efficiency of the separation process (Bahrami et al., 2019).

In this context, the separation method "Goods to Person [GTP]" stands out since the products are brought to the operator through the automation of storage and withdrawal activities with the use of "Automated storage and retrieval systems [AS/RS]" such as stacker cranes and miniloads (Koster et al., 2007).

In addition to reducing travel time, other objectives in the design of a warehouse are related to minimizing total costs (investment and operational) and processing time of an order and maximizing the use of space, equipment, labor and accessibility of products (Koster et al., 2007).

One of the most important issues in warehouse design relates to the choice of storage which includes decisions on the assignment of "Stock Keeping Units [SKUs]" in the various departments, the scheduling of inventory movements between departments, the assignment of SKUs to different zones, and the definition of a storage location within a department/zone (Gu et al., 2007). The three criteria most often used to make these decisions are related to SKU turnover (ABC), inventory need and "Cube per Order Index [COI]" which is defined by the ratio between the required inventory space and SKU turnover (Frazelle, 2002).

The Storage Location Assignment Problem (SLAP) has been addressed in several studies. According to Reyes et al. (2018), the SLAP problem refers to the allocation of products in a storage space aiming at optimizing the costs of material handling and the utilization of storage space. The problem depends on parameters such as the layout, availability and capacity of the storage area, physical characteristics of the products, frequency of receipt and demand profile.

This problem was initially presented by Hausman et al. (1976) where 3 most representative strategies were analyzed: (a) random storage, (b) dedicated storage and (c) storage by product class (grouping of SKUs). Their study identified that a storage by classes

provided a significant increase in the efficiency in the movement of materials when compared to the random approach and showed a more robust practical application when compared to dedicated storage.

Several techniques for different scenarios are presented in the literature to solve the SLAP problem. In their paper, Silva et al. (2020) presents a model that integrates the storage problem with the order picking problem. Fontana et al. (2020) applies a multicriteria model to solve the problem in a manual warehouse.

In their research, Chen et al. (2007) analyzes the problem for an automatic AS/RS system by applying a Tabu Search algorithm. Mirzaei et al. (2021) propose a Cluster-based allocation to improve the performance of an automatic AS/RS system with GTP separation. In his doctoral thesis, Kofler (2015) presents a model for optimizing the allocation problem under dynamic conditions. Lorenc et al. (2021) uses "Artificial Neural Network [ANN]" and Clustering techniques to solve the SLAP problem.

This research will focus on the problem of storage location assignment through the grouping of SKUs and the selection of the best group for allocation in an automatic AS/RS system with GTP separation. First, different unsupervised machine learning techniques will be investigated to identify which ones can help solve the problem and later a comparison will be made between the applied techniques and the techniques usually used in the literature such as ABC and COI. Because it is eminently applied research, the study will go directly to the description of the methodology.

Methodology

The proposed research had an exploratory and descriptive objective since it initially sought to understand which Machine Learning techniques can assist in the grouping of SKUs to later perform a comparative description of the results obtained through the applied techniques.

The nature of the data is of quantitative origin and refers in a general way to the characteristics of the products and the order history of a logistics operation. The database used was obtained from a real logistics operation.

The research design involved the implementation of a Machine Learning algorithm to assist in the problem of grouping SKUs. The code was written in the R language and exploratory multivariate techniques such as Principal Component Factor Analysis and Cluster Analysis, the Autoencoder artificial neural network and the "Analytic Hierarchy Process [AHP]" method were used to compare the results.

The research problem

The operating costs of a warehouse are strongly influenced by the decisions made during the intralogistics project. The development of the project goes through the stages of data analysis, selection of equipment and elaboration of the layout.

Each stage has a great influence on the other and there are several possible solutions making this activity highly complex (Rouwenhorst et al., 2000). The design and operation of a warehouse encompasses several problems and there are several possible technologies to solve them.

Table 1 illustrates the different problems and decisions present in the operation of a warehouse and table 2 presents some of the logistics technologies available in the market.

Table 1 – Description of the problems and operational decisions of a warehouse

Processes	Problems	Decisions
Receiving and shipping	Input and output of materials	Allocation of the order to the truck
		Allocation of the truck to the dock
		Truck loading scheduling
Storage	Allocation of sku to department	Allocation of skus to different warehouse departments
		Space allocation
	Zoning	Sku allocation for zones
		Allocation of separators to zones
	Storage location allocation	Storage location allocation
		Specifying storage classes (for class-based allocation)
Order Separation	Order batch	Batch size
		Allocation of orders to batches
	Routing and sequencing	Routing and sequencing of order routes
		Choice of resting point (for as/rs systems)
	Draw	Assignment of orders for ramps

Source: Adapted from (Gu et al., 2007)

Table 2 – Description of logistics technologies

Processes	Kind	Technology
Storage	Pallet	Pallet truck
		Mobile bases
		Channel storage
		Stacker cranes
	Box	Shelves
		Flowracks
		Miniload
		Shuttle
Order Separation	Man to the product	Guided by rf
		Guided by light
		Voice-guided
	Product to man	Vertical warehouse
		gtp stations
	Automatic	A-frame
		Robots

Source: Adapted from (ssi-schaefer.com)

This research focused on the problems related to the storage process more specifically in the decision to allocate SKUs to zones. According to Gu et al. (2007), the zoning problem is considered a design decision when the zones have different storage technologies. As for the technology, the automatic AS/RS shuttle type system was considered for storage of boxes and separation of fractional units in a GTP station.

This paper sought to answer the following questions:

- Given a set of SKUs N with a known demand X, what is the best group of these SKUs G < N to be allocated in an automatic AS/RS system with C < X capability?
- What unsupervised machine learning techniques can assist in the formation of these groups?
- Are the groups formed through these techniques more suitable than the groupings carried out through conventional techniques such as the ABC and the IOC?

The following premises were adopted:

- Demand based on the operation of fractional units;
- Storage in plastic boxes with dimensions of 600 x 400 x 220 mm, occupancy factor of 70% and capacity of 35kg;
- Only 1 SKU stored per box;
- Need for inventory based on the 30-day inventory policy;
- Separation will be made on request, i.e. without batch formation;

The database

The database used was obtained from the logistics operation of a company in the retail sector. Two CSV files were used: the product register and the order history.

The product register is a table containing the following information of the SKUs:

- SKU (SKU Code)
- Length (mm)
- Width (mm)
- Height (mm)
- Weight (grams)
- Volume (liters)
- Parts per box [PPC]
- Parts per pallet [PPP]

The order history is a table containing the following order information for the period of May and June 2016:

- Order (Order Code)
- Date
- SKU (SKU Code)
- Quantity

The product registration table has 14,424 rows that correspond to the number of SKUs present in the operation and the order history table has 1,078,512 rows corresponding to the order lines that represent the number of distinct SKUs present in the orders.

The first step was to import the tables and, subsequently, join them into a single table through the SKU code. With the consolidated information, the following variables were created:

- Fractionated
- Fractional lines
- Fractional volume

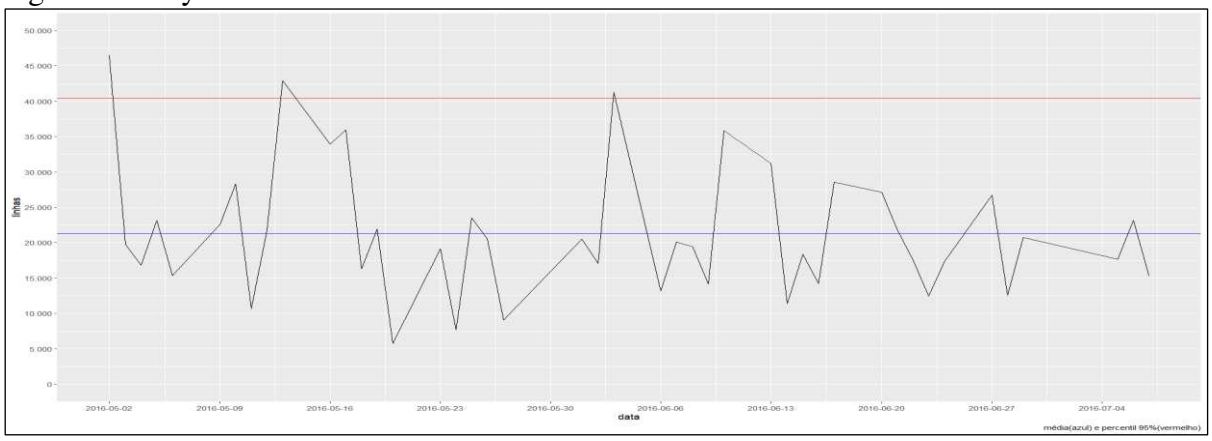
Fractions correspond to the units that are ordered in their smallest unit of measurement. They are calculated by dividing the total quantity by the PPP and the rest of it obtained in the division by the PPP. Fractionate lines are the order lines that contain fractionals, and the fractional volume is obtained by multiplying the volume by the fractional units.

The following example illustrates the calculation performed by considering an order line with a quantity equal to 1,000 for SKU A (PPC = 45 and DPI = 900):

- 1,000 / 900 = 1 with 100 remainder
- 100 / 45 = 2 with rest 10
- The order line contains 1 pallet, 2 boxes and 10 fractions.

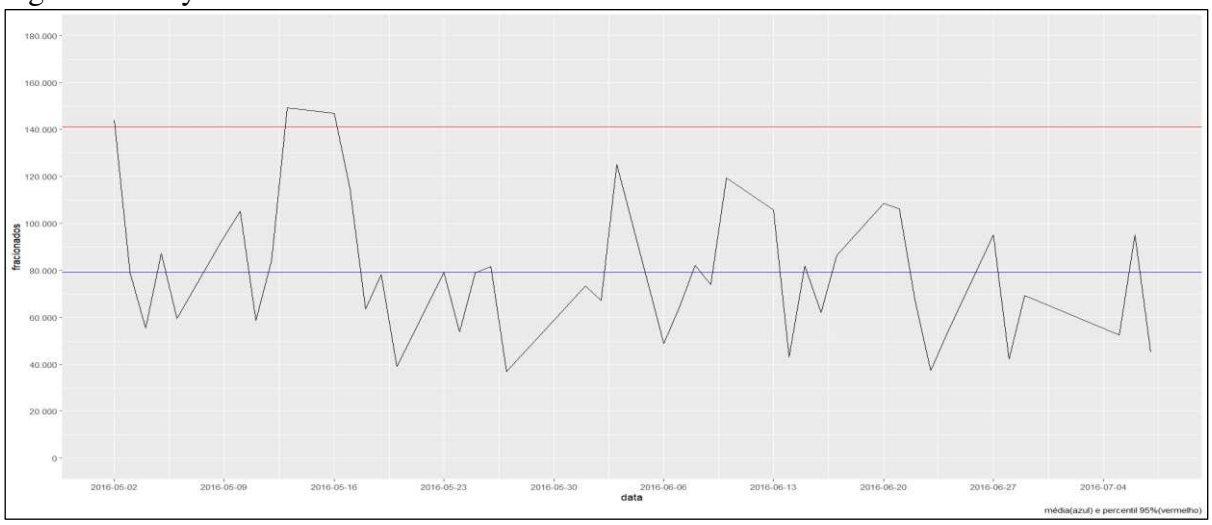
Next, a filter was performed so that only the rows containing fractions were kept in the table. The present research focused only on the operation of fractionated. The demand shipped in boxes and pallets was not considered. Figures 1 and 2 show the demands separated by lines and units, as well as the means and 95% percentiles.

Figure 1: Daily demand in rows



Source: Original research data

Figure 2: Daily demand in units



Source: Original research data

Techniques used

There are three most commonly used criteria for assigning a storage site. In the random allocation the products are stored in random positions in the warehouse and in the dedicated allocation each product is given a specific position. The third type is known as class-based allocation and combines the characteristics of the previous criteria since it divides the products into classes and within these classes the storage occurs randomly (Gu et al., 2007).

This research focused on the class-based criterion we call the grouping of SKUs. In this context, the SLAP problem consists of assigning SKUs to groups and allocating these groups within a storage area (Bahrami et al., 2019).

According to Frazelle (2002), the three most used criteria in the allocation of groups are:

- ABC: Based on the turnover of products that is associated with the amount of storage and withdrawal operations per unit of time.
- Inventory: Based on the required inventory space of the product group.
- COI: Index defined by the ratio between inventory space and product turnover.

Additionally we can cite the XYZ criterion described in Stojanović et al. (2017):

- XYZ: Based on demand variability, this criterion uses the coefficient of variation [CV], which is defined by the ratio of the standard deviation to the mean.

The 4 criteria presented were considered for the grouping of SKUs and in the selection of groups for the proposed storage system. Subsequently, unsupervised machine learning techniques were used to assist in solving the problem.

Exploratory multivariate techniques allow studying the relationship between variables in a database, investigating the correlation between variables, elaborating a ranking of observations, and grouping observations and variables (Fávero and Belfiore, 2017). The main techniques for metric variables are:

Principal Component Factor Analysis: A technique that uses correlation coefficients to group variables and generate factors (Fávero and Belfiore, 2017). This technique was used in the research to elaborate a ranking based on the factors. This ranking was used to group the SKUs. The factors generated in this technique were also used as a basis for cluster analysis.

Cluster Analysis: Technique that allows to verify the existence of similar behaviors between the observations for the creation of clusters with internal homogeneity. The groupings are divided into the hierarchical and the non-hierarchical (Fávero and Belfiore, 2017). The non-hierarchical k-means technique was used in this research to assist in the problem of grouping SKUs.

Additionally, the Autoencoder technique was used. According to Boehmke (2019), Autoencoder is a neural network that is trained to learn efficient representations of the input data. Among other applications, this technique can be used to reduce the dimensionality of the database. When Autoencoder uses only linear activation functions, it has very similar results to "Principal Components Analysis [PCA]".

Finally, the AHP technique was used to perform a comparison of the results obtained. AHP is a multicriteria methodology that aims to select the best alternatives through a process that considers different evaluation criteria (Santos et al., 2021).

Results and Discussion

Initially, a new table (SKU table) was created gathering the following dimensional and movement information from the SKUs:

- Fractionated
- Lines
- Volume
- Weight
- Days (corresponds to the number of days the SKU was moved)
- CV lines (coefficient of variation of fractional lines over days)
- Fractional CV (coefficient of variation of fractionated over days)
- CV volume (coefficient of variation of the volume of fractionate in the lines)
- Boxes (number of boxes needed in stock)
- COI (index obtained by dividing the sum of fractions by boxes)

To calculate the variable Boxes, the following assumptions were considered:

- Box with dimensions of 600 x 400 x 220 mm
- 70% occupancy factor
- Net volume of 37 litres
- Maximum weight of 35 kg
- 30-day inventory policy

The SKU table was used as the basis for the analyses performed in this research. Table 3 shows that the fractional measurements, lines and caixas_estoque present a standard deviation

about 3 times higher than the mean, indicating a great dispersion of observations for these measurements. The other measures, in turn, exhibit a more uniform behavior.

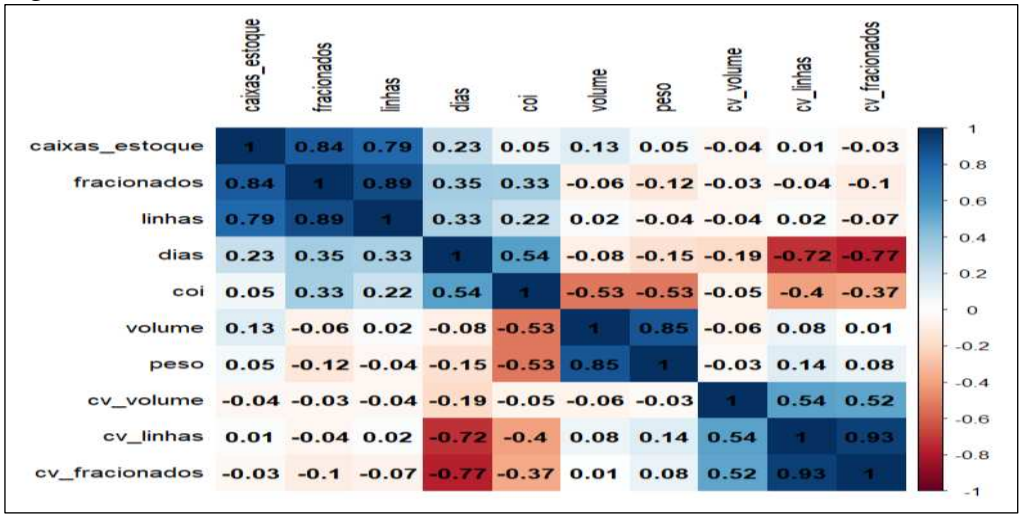
Table 3: Descriptive statistical analysis

Measure	Average	Standard deviation	1% percentile	25% percentile	50% percentile	75% percentile	99% percentile
fractionated	6,51	17,87	0,02	0,64	2,09	5,27	92,70
Lines	1,75	4,79	0,02	0,23	0,59	1,30	28,50
cv_linhas	2,23	1,50	0,66	1,21	1,69	2,72	6,63
cv_fracionados	2,70	1,45	0,85	1,67	2,26	3,30	6,63
cv_volume	1,45	1,90	0,00	0,79	1,06	1,38	10,00
Days	17,64	11,65	1,00	8,00	17,00	26,00	43,00
volume	0,62	0,31	0,11	0,40	0,66	0,74	1,59
weight	0,36	0,19	0,06	0,20	0,36	0,48	0,87
caixas_estoque	3,61	11,59	1,00	1,00	1,00	3,00	42,97
see	1,55	1,51	0,02	0,61	1,23	1,84	7,80

Source: Original research results

Figure 3 shows the formation of groups of variables according to the correlation between them. The variables caixas_estoque, fractions and lines have a high correlation with each other. The same occurs with variables cv_linhas and cv_fracionados which also correlate with the variable days. Volume and weight form a third group with high correlation.

Figure 3: Correlation matrix



Source: Original research results

Univariate analysis

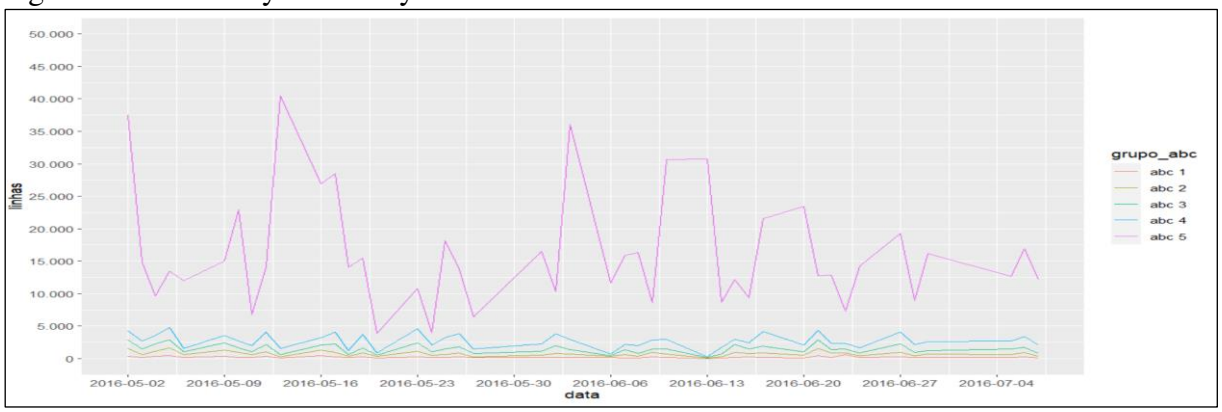
From the SKU table, a univariate analysis was initially performed. For each technique, the SKUs were ordered in ascending order and divided into 5 groups with the same number of observations according to the following variables:

- ABC: lines
- Stock: caixas_estoque
- IOC: IOC
- XYZ: cv_linhas

The results are presented in the following charts.

Figure 4 shows that the abc 5 group presents the highest demand for lines, since it gathers 20% of the SKUs with the highest value for this variable.

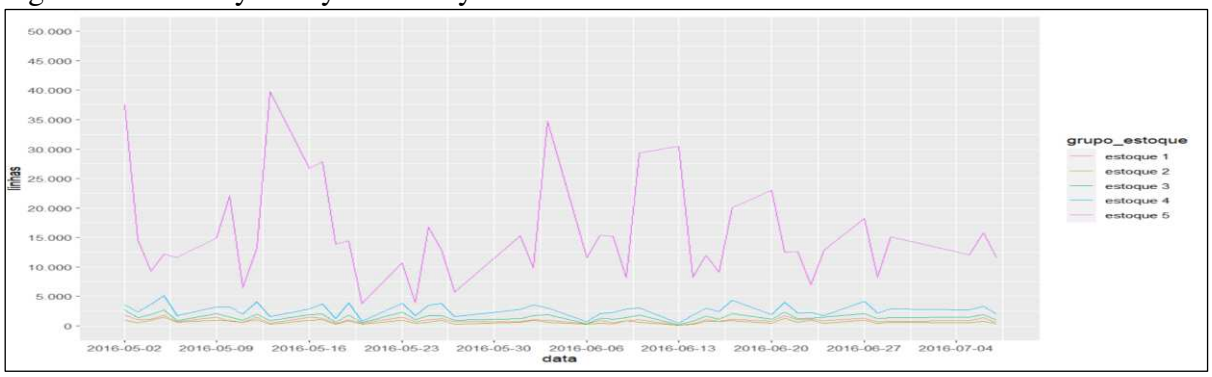
Figure 4: ABC Analysis – Daily demand



Source: Original research results

Figure 5 shows that the stock group 5 accounts for the highest demand in lines. This fact was expected because these variables have a high correlation.

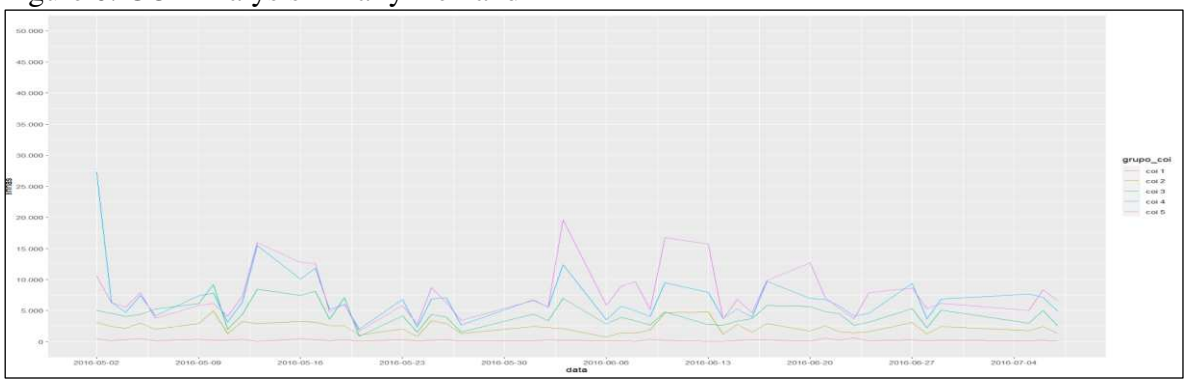
Figure 5: Inventory Analysis – Daily Demand



Source: Original research results

Figure 6 shows a different behavior from that previously observed. The demand between the groups was better divided without presenting a marked concentration in the coi group 5. This is due to the fact that the variThe coi does not present a high correlation with the variable lines.

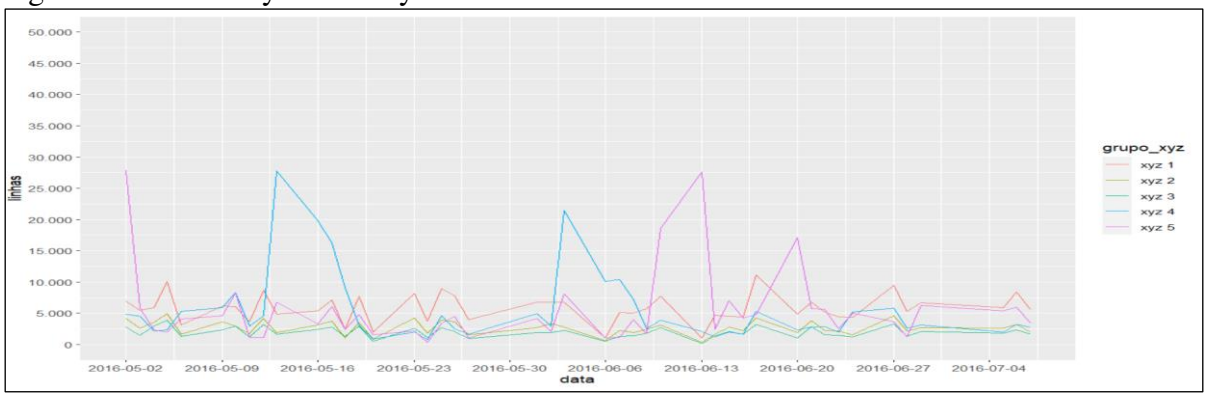
Figure 6: COI Analysis – Daily Demand



Source: Original research results

Finally, Figure 7 highlights the fact that, for this variable, the group xyz 5 represents the 20% of SKUs with the greatest variability in line demand per day, while group xyz 1 has a more uniform behavior.

Figure 7: XYZ Analysis – Daily Demand



Source: Original research results

As we can see in the graphs, the groups formed have very different characteristics. Table 4 presents the characteristics of the clusters formed. The best group is the one that can:

- maximize the amount of skus,
- maximize the average demand of lines per day (linhas_média),
- maximise the ratio of SKUs per box (sku_caixa),
- maximize the ratio of average demand to maximum lines (fator_linha),
- maximise the ratio of lines per box (linha_caixa),
- minimize the need for inventory (boxes),
- minimize the maximum demand for lines per day (linhas_max).

As expected, the abc 5 group has the highest demand for lines, while the stock group 5 has the highest demand for boxes. The group coi 5 has the best ratio of lines per box and group xyz 1 is, among the groups with more than 5,000 lines, the one with the best ratio between middle and maximum line.

A first comparison can be made between the groups xyz 1, xyz 4 and xyz 5. The groups have the same amount of SKUs, a similar inventory need, and average line demand. However, the maximum demand for lines in groups 4 and 5 is more than double the demand for group 1. Among these, group 1 would be the most appropriate because, by having a more uniform demand, it requires less performance and, consequently, less investment to deliver the same result.

Each group stood out in some metric according to the grouping technique used, but no group presented a good balance considering the set of metrics.

Table 4: Univariate analysis – Summary table

group	SKUs	Boxes	linhas_média	linhas_max	sku_caixa	fator_linha	linha_caixa
abc 1	2.441	2.484	198	584	0,98	0,34	0,08
abc 2	2.441	2.661	737	1.673	0,92	0,44	0,28
abc 3	2.441	3.843	1.476	2.893	0,64	0,51	0,38
abc 4	2.441	6.209	2.716	4.794	0,39	0,57	0,44
abc 5	2.440	28.858	16.221	40.519	0,08	0,40	0,56
xyz 1	2.441	11.492	5.805	11.108	0,21	0,52	0,51
xyz 2	2.441	5.881	2.675	4.969	0,42	0,54	0,45

xyz 3	2.441	4.666	1.939	3.899	0,52	0,50	0,42
xyz 4	2.441	10.654	5.436	27.753	0,23	0,20	0,51
xyz 5	2.440	11.362	5.493	27.854	0,21	0,20	0,48
Whistle 1	2.441	2.450	244	642	1,00	0,38	0,10
Whistle 2	2.441	8.095	2.327	4.961	0,30	0,47	0,29
Whistle 3	2.441	11.076	4.329	9.177	0,22	0,47	0,39
Whistle 4	2.441	12.680	6.861	27.277	0,19	0,25	0,54
Whistle 5	2.440	9.754	7.588	19.595	0,25	0,39	0,78
Stock 1	2.441	2.441	942	1.833	1,00	0,51	0,39
Stock 2	2.441	2.441	656	1.446	1,00	0,45	0,27
Stock 3	2.441	3.435	1.462	2.682	0,71	0,55	0,43
Stock 4	2.441	5.866	2.688	5.105	0,42	0,53	0,46
Stock 5	2.440	29.872	15.600	39.785	0,08	0,39	0,52

Source: Original research results

Multivariate analysis

Following the study, multivariate techniques were considered, that is, techniques that use two or more variables and also consider the relationship between them.

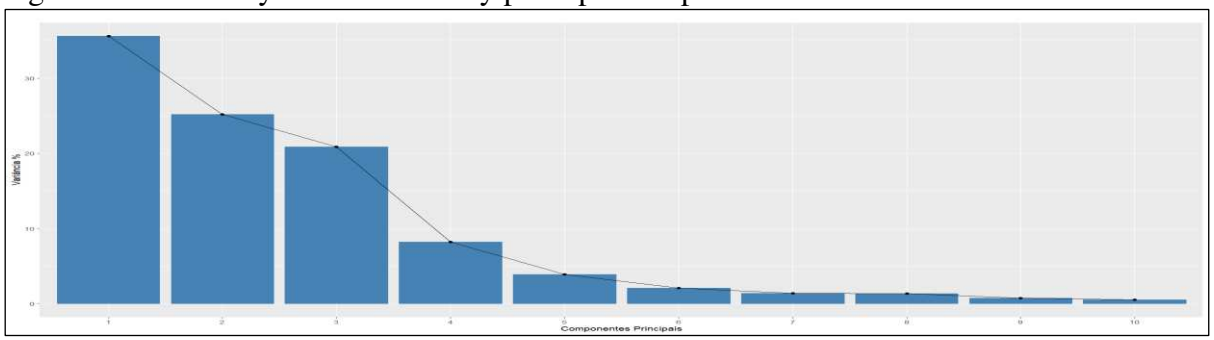
PCA Technique

Initially, the SKU table was standardized through the scale command. We then obtained the matrix of correlations in which the following tests were applied to verify the global adequacy of the extraction of factors:

- Kaiser-Meyer-Olkin (KMO) statistic: the result obtained was 0.7 which is considered reasonable (Fávero and Belfiore, 2017).
- Bartlett's sphericity test: the P-value obtained was 3.92-154 proving that Pearson's correlations between the pairs of variables are statistically different from 0 and, therefore, factor analysis is appropriate (Fávero and Belfiore, 2017).

Once the adequacy was verified, the factor analysis was performed using the prcomp function of the stats library of the R language. To determine the number of factors we can use the latent root criterion that considers only the factors that have eigenvalues greater than 1 (Fávero and Belfiore, 2017). The first 3 main components meet this criterion and were considered in this research.

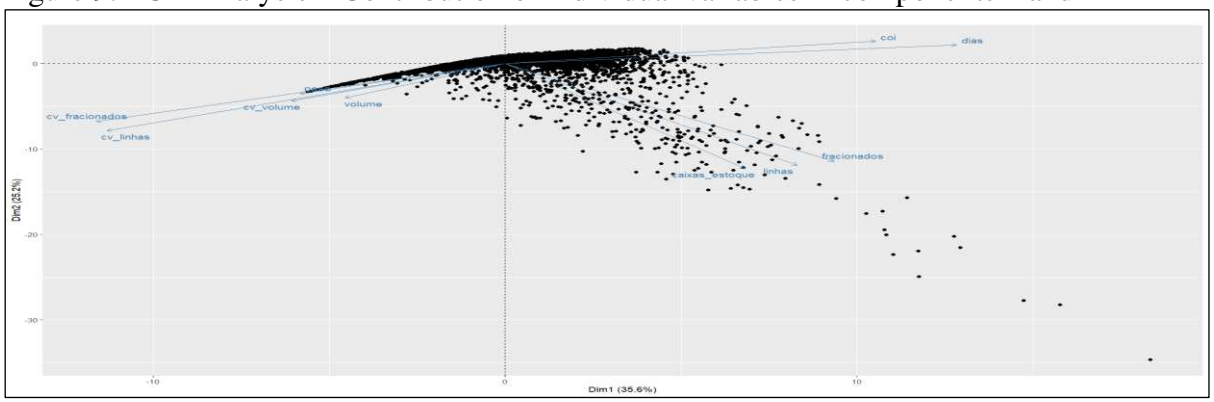
Figure 8: PCA Analysis – Variance by principal components



Source: Original research results

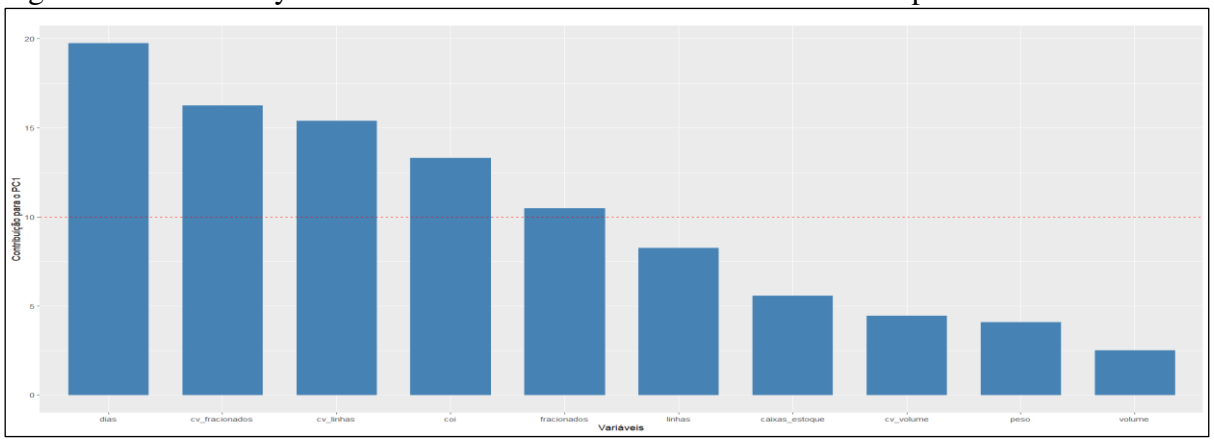
The first two components account for 60% of the variance. Figures 13 and 14 illustrate the contribution of individual variables in the formation of components.

Figure 9: PCA Analysis – Contribution of individual variables in components 1 and 2



Source: Original research results

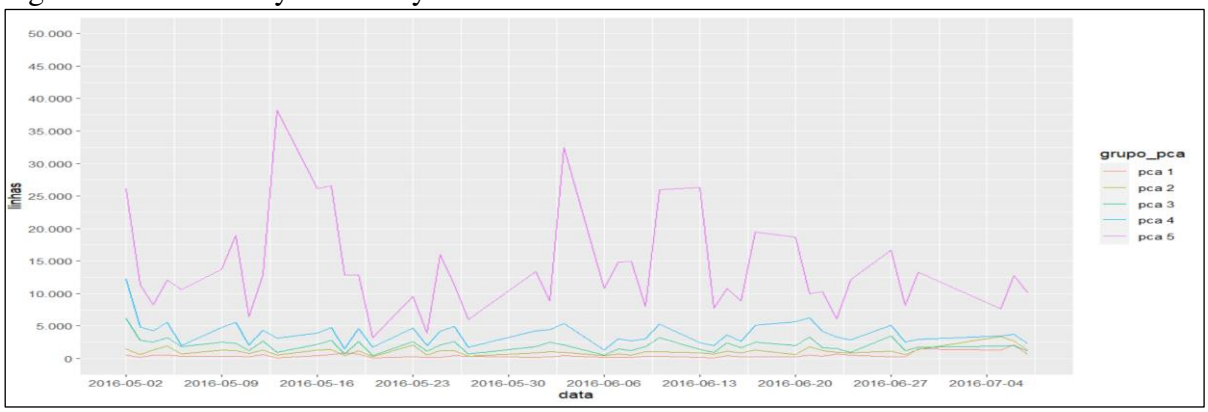
Figure 10: PCA Analysis – Contribution of individual variables in component 1



Source: Original research results

Subsequently, the SKUs were ordered and grouped according to the first main component. The results are shown in Figure 11 and Table 5.

Figure 11: PCA Analysis – Daily Demand



Source: Original research results

Table 5 shows that the groups formed present behavior similar to those formed by the ABC and Stock techniques. This means that although the first main component Load the variance of several variables, the groups formed still do not have a good balance considering the set of metrics.

Table 5: PCA Analysis – Summary Table

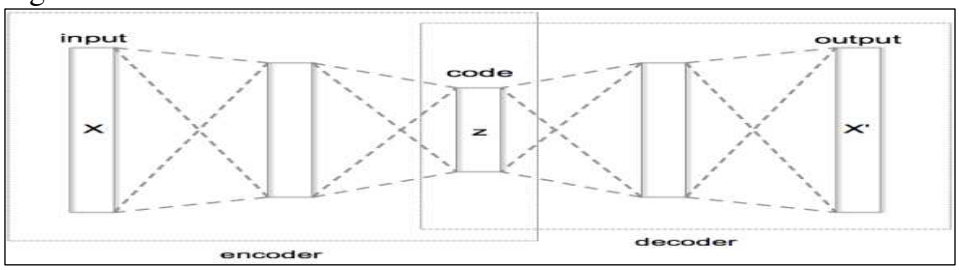
group	SKUs	Boxes	linhas_média	linhas_max	sku_caixa	fator_linha	linha_caixa
pca 1	2.441	2.748	424	2.054	0,89	0,21	0,15
pca 2	2.441	3.421	1.090	3.340	0,71	0,33	0,32
pca 3	2.441	4.875	2.002	6.122	0,50	0,33	0,41
pca 4	2.441	7.498	3.856	12.209	0,33	0,32	0,51
pca 5	2.440	25.513	13.976	38.259	0,10	0,37	0,55

Source: Original research results

Autoencoder Technique

According to Boehmke (2019), the Autoenconder has a structure similar to a direct neural network. The main difference is that when used in an unsupervised context, the number of neurons in the output layer (X') is equal to the number in the input (X).

Figure 12: Autoencoder Structure



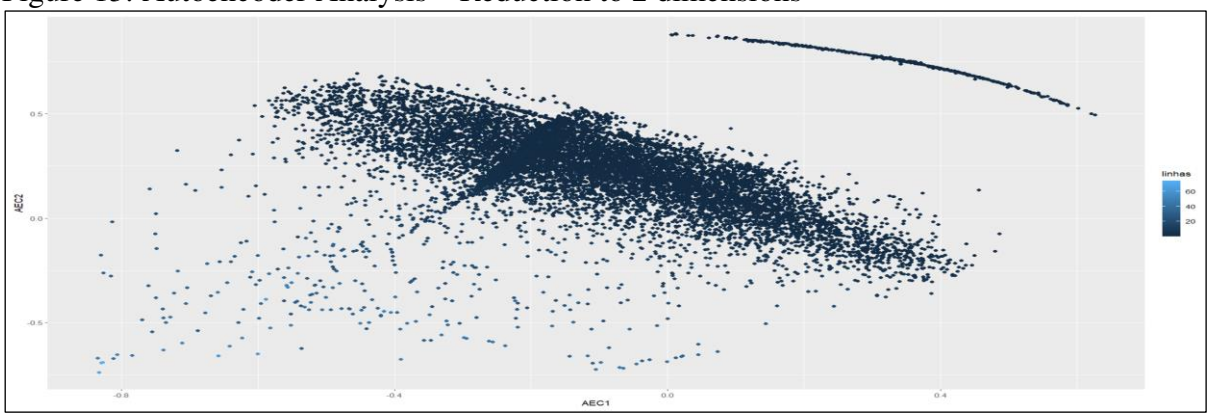
Source: Boehmke (2019)

In reducing dimensionality the goal is to create a set of codes (z) that adequately represent the input data (Boehmke, 2019).

To apply the Autoenconder technique, the keras library was used. The model considered has structure with 3 hidden layers, two with 7 units and the innermost was tested with 2 and 3 units.

Figure 13 shows the reduction of the base with 10 variables for 2 dimensions. In the reduction to 2 dimensions, a "Mean squared error [MSE]" of 0.0938 was found.

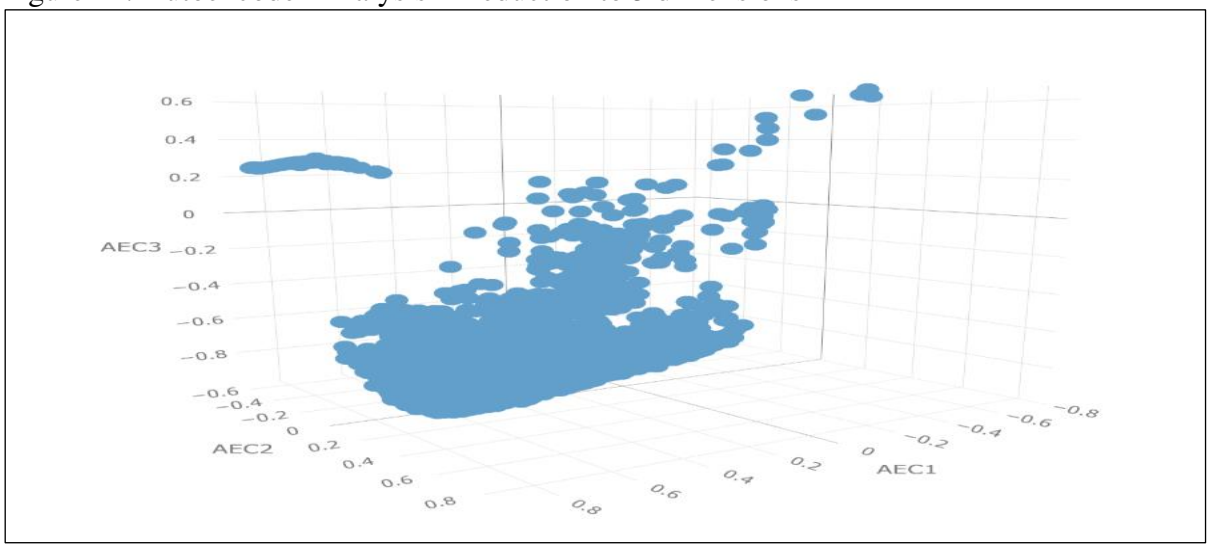
Figure 13: Autoencoder Analysis – Reduction to 2 dimensions



Source: Original research results

Figure 14 shows the reduction of the base with 10 variables for 3 dimensions. In the reduction to 3 dimensions, an MSE of 0.0636 was found.

Figure 14: Autoencoder Analysis – Reduction to 3 dimensions



Source: Original research results

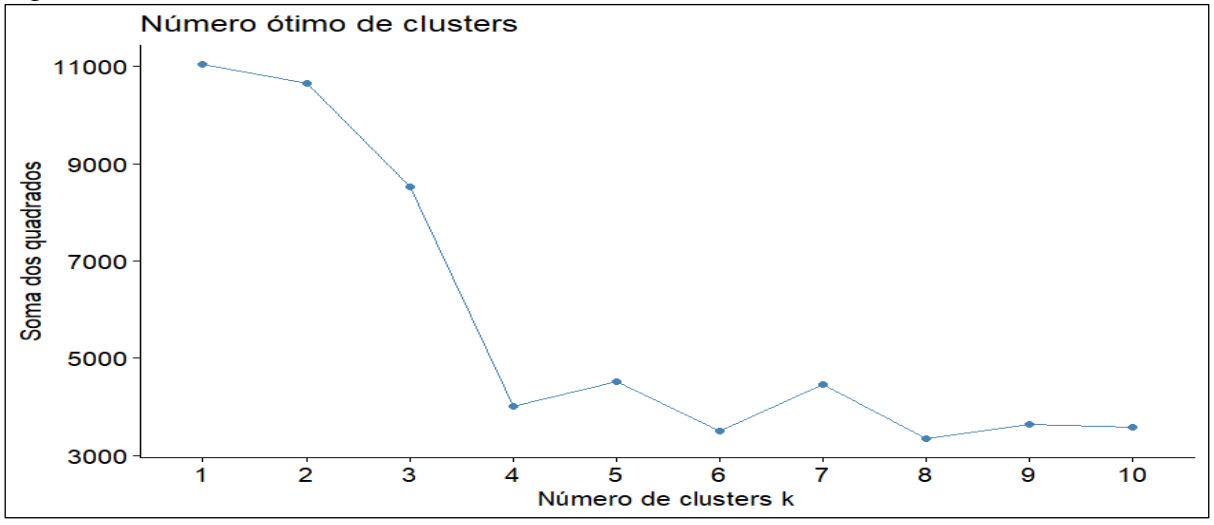
Clustering Technique

In the application of the Clustering technique, the kmeans function of the stats library was used. The technique was performed considering the following scenarios:

- PCA: 2 main components
- PCA: 3 main components
- Autoencoder: represented in 2 dimensions
- Autoencoder: represented in 3 dimensions

The determination of the number of clusters was performed through the elbow method applied to the first 2 scenarios. Figure 15 shows an ideal value of 4 clusters.

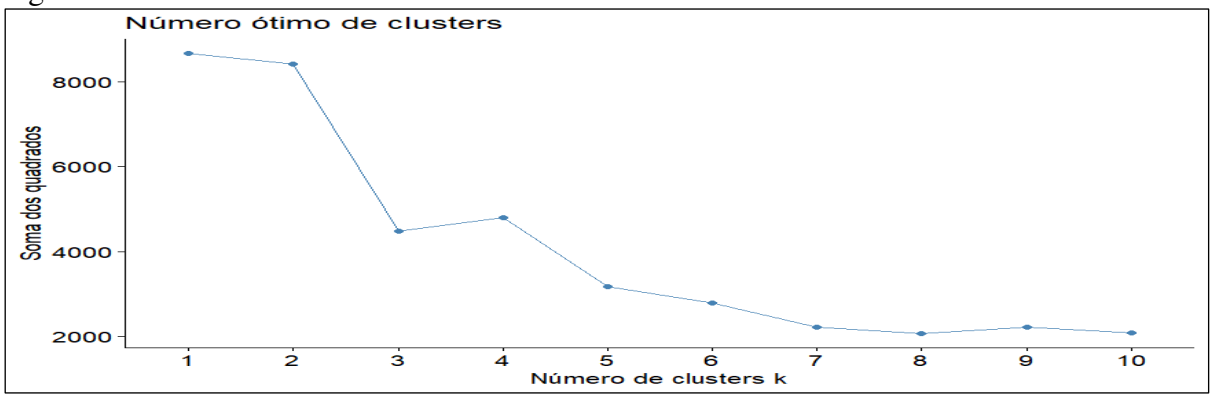
Figure 15: Elbow method – PCA with 2 factors



Source: Original research results

In Figure 16 it was verified that the ideal value would be 5 clusters. The present research considered the formation of 3, 4 and 5 clusters for each of the proposed scenarios.

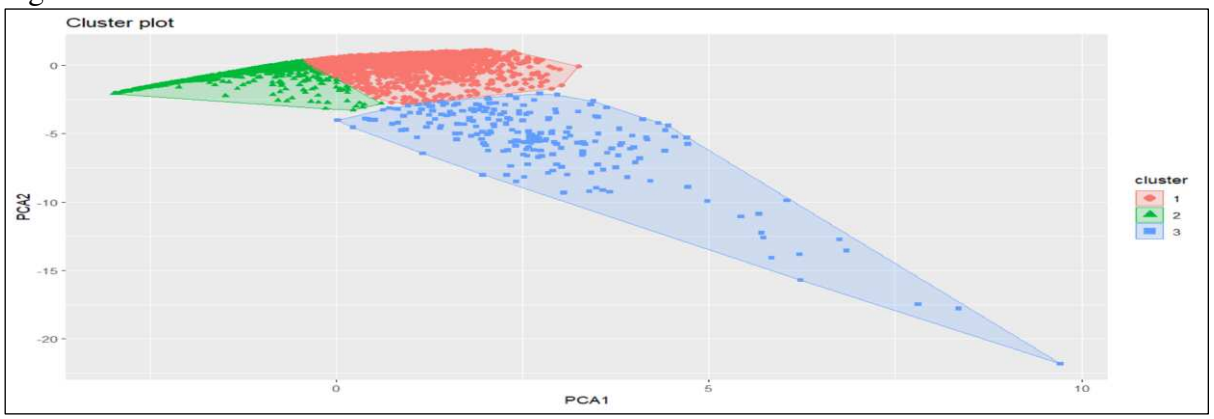
Figure 16: Elbow method – PCA with 3 factors



Source: Original research results

The following graphs present each of the scenarios considering the formation of 3 clusters. Details about the clusters of 4 and 5 clusters can be found in table 6.

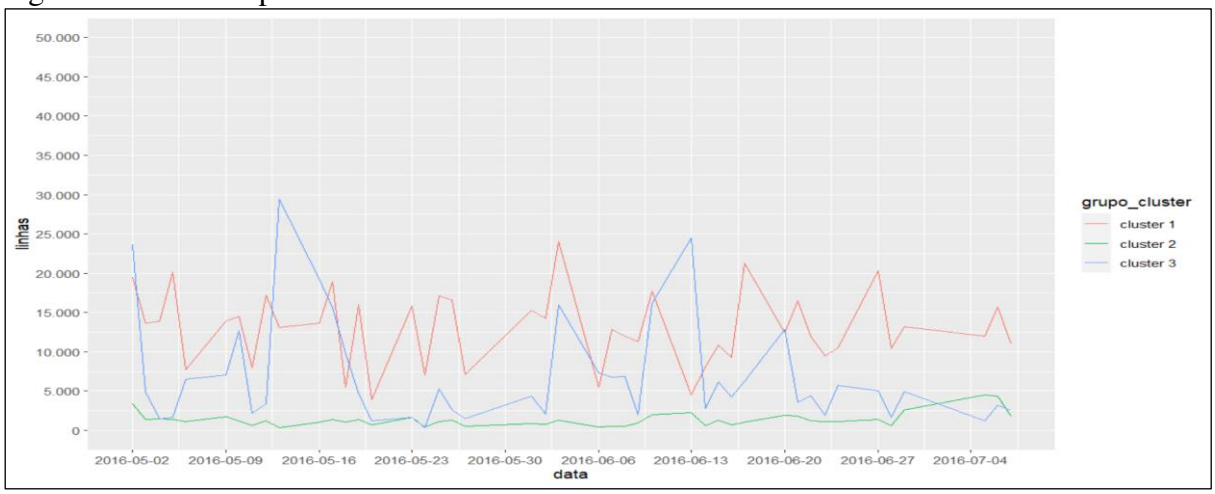
Figure 17: Clusters formed – PCA with 2 factors



Source: Original research results

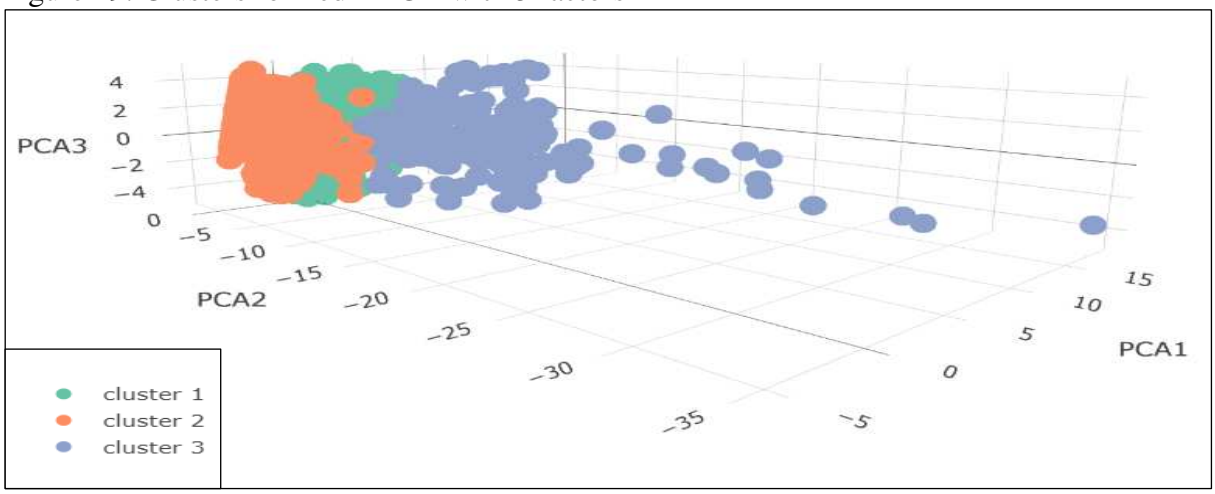
In Figure 18 Note that cluster 2 gathers SKUs with low demand while cluster 3 contains SKUs with high demand with the presence of daily peaks. Cluster 1 has high and uniform demand.

Figure 18: Demand per cluster – PCA with 2 factors



Source: Original research results

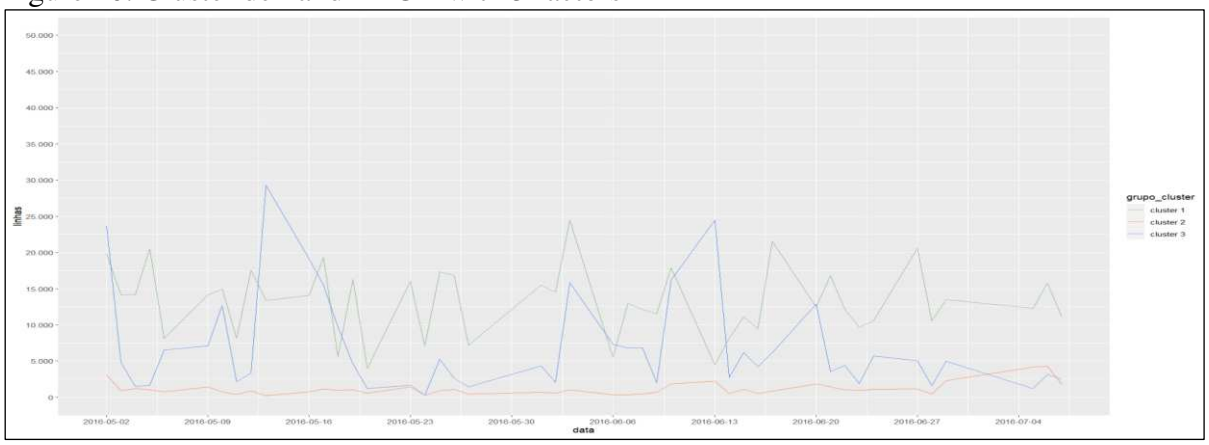
Figure 19: Clusters formed – PCA with 3 factors



Source: Original research results

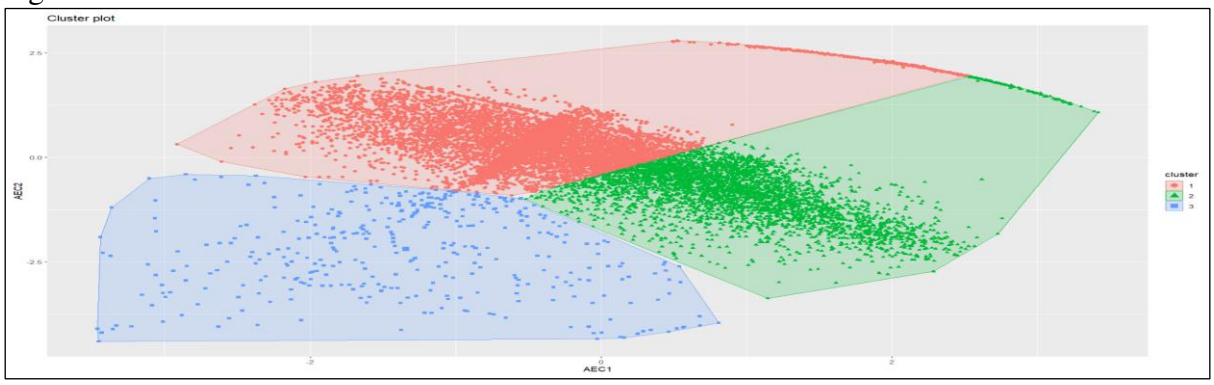
Figure 20 shows the similarity of the results produced between the PCA technique applied considering 2 and 3 factors.

Figure 20: Cluster demand – PCA with 3 factors



Source: Original research results

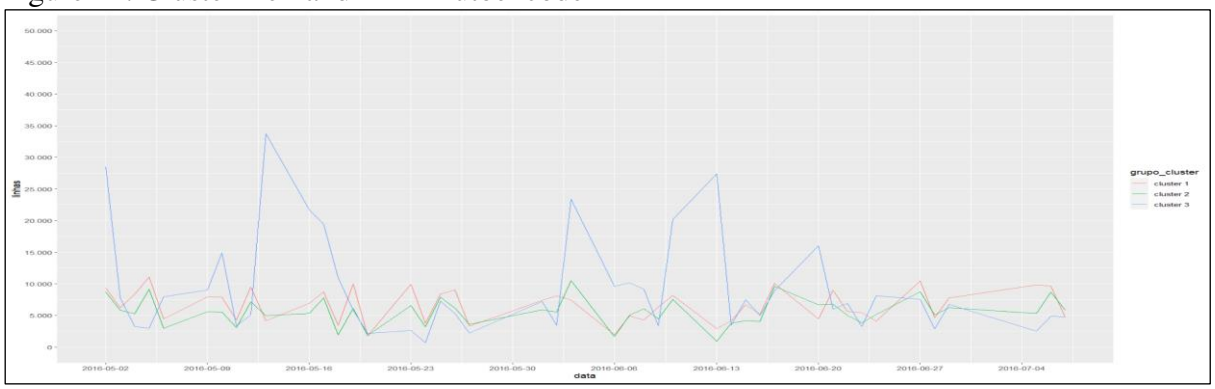
Figure 21: Clusters formed – Autoencoder in 2D



Source: Original research results

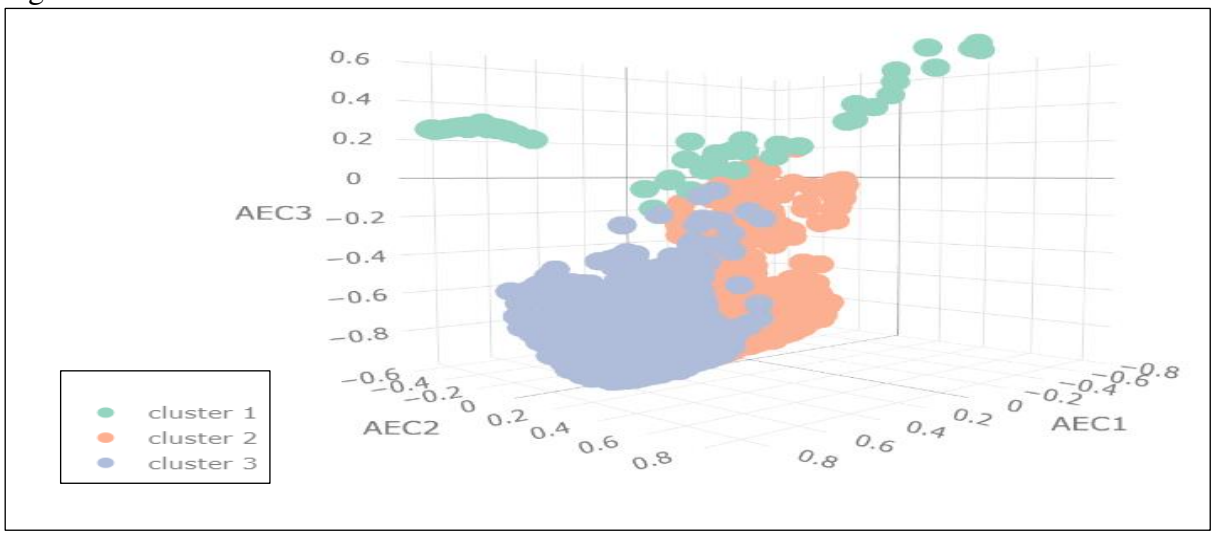
Figure 22 shows that cluster 3 concentrates SKUs with high demand and variation. Clusters 1 and 2, in turn, behave much like smaller, more uniform demand.

Figure 22: Cluster Demand – 2D Autoencoder



Source: Original research results

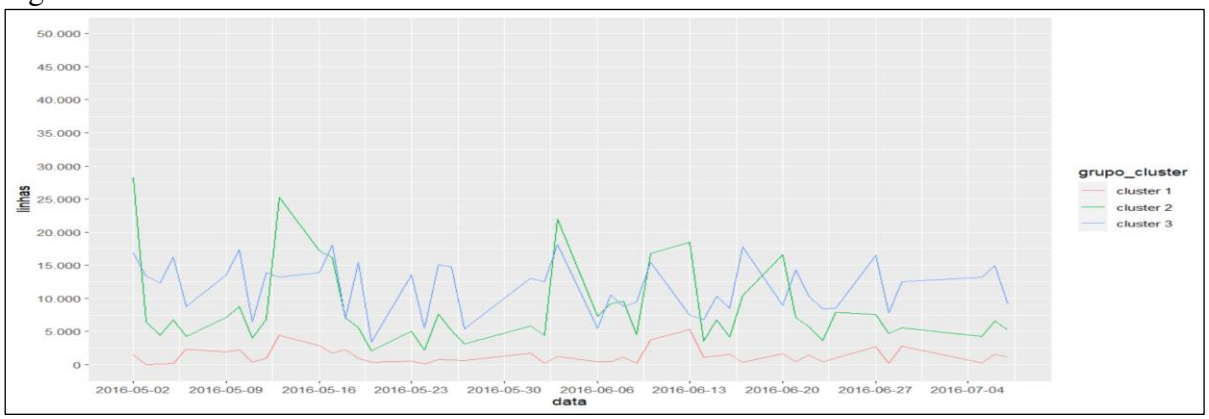
Figure 23: Clusters formed – 3D Autoencoder



Source: Original research results

In Figure 24 it is noted that the reduction to 3 dimensions of the Autoencoder produced very different results from the 2D reduction, but quite similar to those obtained with the PCA technique.

Figure 24: Cluster Demand – 3D Autoencoder



Source: Original research results

Table 6 shows that the groups formed through the Clustering technique have very different characteristics. Initially analyzing the 3 clusters formed from the first 2 main components we verified that group 3 gathers few SKUs with high demand in lines and boxes. The ratio between midline and maximum line is also the worst of the 3 groups. Group 2, in turn, contains SKUs with low demand and has the best ratio of boxes per SKU. Finally, group 1 is composed of many SKUs that have an average and more uniform demand, making the line factor the largest among the 3. The same behavior is repeated for the other groups formed through the PCA and the Autoencoder.

Table 6: Clustering Analysis – Summary Table

group	Skus	Boxes	linhas_média	linhas_max	sku_caixa	fator_linha	linha_caixa
PCA2D 1-3	8.232	24.880	13.028	24.045	0,33	0,54	0,52
PCA2D 2-3	3.752	5.201	1.351	4.480	0,72	0,30	0,26
PCA2D 3-3	220	13.974	6.969	29.427	0,02	0,24	0,50
PCA2D 1-4	4.419	17.916	10.213	20.385	0,25	0,50	0,57
PCA2D 2-4	1.661	2.237	523	3.460	0,74	0,15	0,23
PCA2D 3-4	216	13.859	6.896	29.207	0,02	0,24	0,50
PCA2D 4-4	5.908	10.043	3.717	7.377	0,59	0,50	0,37
PCA2D 1-5	706	853	147	2.448	0,83	0,06	0,17
PCA2D 2-5	5.494	11.418	4.815	8.648	0,48	0,56	0,42
PCA2D 3-5	2.551	13.308	8.213	18.116	0,19	0,45	0,62
PCA2D 4-5	209	13.684	6.738	28.508	0,02	0,24	0,49
PCA2D 5-5	3.244	4.792	1.436	4.175	0,68	0,34	0,30
PCA3D 1-3	8.625	25.593	13.253	24.443	0,34	0,54	0,52
PCA3D 2-3	3.360	4.520	1.147	4.255	0,74	0,27	0,25
PCA3D 3-3	219	13.942	6.948	29.240	0,02	0,24	0,50
PCA3D 1-4	3.776	11.006	7.028	14.807	0,34	0,47	0,64
PCA3D 2-4	2.066	2.577	586	3.867	0,80	0,15	0,23
PCA3D 3-4	6.146	16.625	6.847	10.923	0,37	0,63	0,41
PCA3D 4-4	216	13.847	6.888	28.896	0,02	0,24	0,50
PCA3D 1-5	216	13.847	6.888	28.896	0,02	0,24	0,50
PCA3D 2-5	676	809	158	2.466	0,84	0,06	0,20
PCA3D 3-5	3.475	4.489	1.320	3.584	0,77	0,37	0,29
PCA3D 4-5	4.557	15.466	6.728	10.735	0,29	0,63	0,43
PCA3D 5-5	3.280	9.444	6.255	13.412	0,35	0,47	0,66
AEC2D 1-3	7.918	17.821	6.623	11.102	0,44	0,60	0,37
AEC2D 2-3	3.897	9.074	5.567	10.516	0,43	0,53	0,61
AEC2D 3-3	389	17.160	9.158	33.724	0,02	0,27	0,53
AEC2D 1-4	3.780	8.991	5.585	10.431	0,42	0,54	0,62
AEC2D 2-4	7.535	18.016	6.979	11.405	0,42	0,61	0,39
AEC2D 3-4	350	16.505	8.772	33.098	0,02	0,27	0,53
AEC2D 4-4	539	543	12	170	0,99	0,07	0,02
AEC2D 1-5	3.791	8.545	4.038	6.963	0,44	0,58	0,47
AEC2D 2-5	328	16.088	8.520	32.841	0,02	0,26	0,53
AEC2D 3-5	1.906	5.269	3.757	8.491	0,36	0,44	0,71
AEC2D 4-5	539	543	12	170	0,99	0,07	0,02
AEC2D 5-5	5.640	13.610	5.021	7.940	0,41	0,63	0,37
AEC3D 1-3	569	5.698	1.277	5.281	0,10	0,24	0,22
AEC3D 2-3	1.844	11.060	8.421	28.206	0,17	0,30	0,76
AEC3D 3-3	9.791	27.297	11.650	18.130	0,36	0,64	0,43
AEC3D 1-4	722	3.912	3.762	12.058	0,18	0,31	0,96
AEC3D 2-4	3.111	14.980	8.915	30.720	0,21	0,29	0,60
AEC3D 3-4	7.802	19.407	7.397	11.788	0,40	0,63	0,38

AEC3D 4-4	569	5.756	1.274	5.281	0,10	0,24	0,22
AEC3D 1-5	567	5.567	1.214	5.281	0,10	0,23	0,22
AEC3D 2-5	2.672	12.353	7.572	23.945	0,22	0,32	0,61
AEC3D 3-5	680	3.870	3.716	12.349	0,18	0,30	0,96
AEC3D 4-5	2.430	5.390	1.807	8.749	0,45	0,21	0,34
AEC3D 5-5	5.855	16.875	7.039	12.090	0,35	0,58	0,42

Source: Original research results

AHP Technique – Comparison of the clusters formed

To perform the comparison and selection of clusters, the AHP method was used. AHP is one of the most well-known decision-making tools and assists in the construction of decision models for a finite number of alternatives. This method is based on peer-to-peer comparison of criteria for the construction of a priority vector that represents the weight of each criterion in the decision process (Santos, 2021).

In his article, Santos (2021) presents a variation of the method known as Gaussian AHP. In this variant the priority vector is constructed from the Gaussian factor which is calculated by the ratio of the standard deviation to the mean for each criterion.

In this research we considered the use of Gaussian AHP to compare and select the best clusters formed. The clusters were divided according to the technique used: PCA2D, PCA3D, AEC2D and AEC3D. The criteria that have to be maximized were those selected for the application of the method.

Table 7 shows that the clusters selected were 1-3, 1-4 and 4-4 for the PCA technique with 2 factors. Table 8 shows the formation of the Gaussian factor and the weight assigned to each parameter.

Table 7: Gaussian AHP – PCA with 2 factors

group	skus	linhas_média	sku_caixa	fator_linha	linha_caixa	AHP-G	RANK
PCA2D 1-3	0,225	0,203	0,068	0,131	0,106	0,156	1
PCA2D 2-3	0,102	0,021	0,149	0,073	0,052	0,084	6
PCA2D 3-3	0,006	0,109	0,003	0,057	0,101	0,048	10
PCA2D 1-4	0,121	0,159	0,051	0,122	0,115	0,113	2
PCA2D 2-4	0,045	0,008	0,153	0,037	0,047	0,061	8
PCA2D 3-4	0,006	0,108	0,003	0,057	0,100	0,047	11
PCA2D 4-4	0,161	0,058	0,121	0,122	0,075	0,112	3
PCA2D 1-5	0,019	0,002	0,171	0,015	0,035	0,052	9
PCA2D 2-5	0,150	0,075	0,099	0,135	0,085	0,111	4
PCA2D 3-5	0,070	0,128	0,039	0,110	0,125	0,088	5
PCA2D 4-5	0,006	0,105	0,003	0,057	0,099	0,047	12
PCA2D 5-5	0,089	0,022	0,139	0,083	0,060	0,081	7

Source: Original research results

Table 8: Formation of Gaussian factor – PCA with 2 factors

Parameter	skus	linhas_média	sku_caixa	fator_linha	linha_caixa
Average	0,083	0,083	0,083	0,083	0,083
Standard deviation	0,071	0,064	0,063	0,040	0,029
Gaussian factor	0,857	0,763	0,759	0,481	0,348
Factor G. Norm.	0,267	0,238	0,237	0,150	0,109

Source: Original research results

The other subsets were also submitted to the method and the 3 best clusters of each make up the set for final comparison. The 12 selected clusters are presented in table 9. Note that the selected clusters have similar characteristics.

Table 9: Selected clusters

group	skus	Boxes	linhas_média	linhas_max	sku_caixa	fator_linha	linha_caixa
PCA2D 1-3	8.232	24.880	13.028	24.045	0,33	0,54	0,52
PCA2D 1-4	4.419	17.916	10.213	20.385	0,25	0,50	0,57
PCA2D 4-4	5.908	10.043	3.717	7.377	0,59	0,50	0,37
PCA3D 1-3	8.625	25.593	13.253	24.443	0,34	0,54	0,52
PCA3D 3-4	6.146	16.625	6.847	10.923	0,37	0,63	0,41
PCA3D 4-5	4.557	15.466	6.728	10.735	0,29	0,63	0,43
AEC2D 1-3	7.918	17.821	6.623	11.102	0,44	0,60	0,37
AEC2D 2-4	7.535	18.016	6.979	11.405	0,42	0,61	0,39
AEC2D 5-5	5.640	13.610	5.021	7.940	0,41	0,63	0,37
AEC3D 3-3	9.791	27.297	11.650	18.130	0,36	0,64	0,43
AEC3D 3-4	7.802	19.407	7.397	11.788	0,40	0,63	0,38
AEC3D 5-5	5.855	16.875	7.039	12.090	0,35	0,58	0,42

Source: Original research results

Table 10 shows that the best group was formed through the PCA with 3 factors and 3 clusters followed by the PCA with 2 factors and 3 clusters and in third was the Autoencoder with 3 factors and 3 clusters. The grouping in 3 clusters had the best result while the grouping in 5 clusters presented the worst score. Considering the mean Gaussian factor, the AEC3D set presented the best mean followed by PCA2D, PCA3D and AEC2D.

Table 10: Gaussian AHP – Final selection

group	skus	linhas_média	sku_caixa	fator_linha	linha_caixa	AHP-G	RANK
PCA2D 1-3	0,100	0,132	0,073	0,077	0,101	0,104	2
PCA2D 1-4	0,054	0,104	0,054	0,071	0,110	0,081	7
PCA2D 4-4	0,072	0,038	0,129	0,072	0,071	0,072	10
PCA3D 1-3	0,105	0,135	0,074	0,077	0,100	0,106	1
PCA3D 3-4	0,075	0,070	0,081	0,089	0,079	0,076	8
PCA3D 4-5	0,055	0,068	0,065	0,089	0,084	0,069	12
AEC2D 1-3	0,096	0,067	0,098	0,085	0,072	0,082	5
AEC2D 2-4	0,091	0,071	0,092	0,087	0,075	0,082	6
AEC2D 5-5	0,068	0,051	0,091	0,090	0,071	0,069	11
AEC3D 3-3	0,119	0,118	0,079	0,091	0,082	0,103	3
AEC3D 3-4	0,095	0,075	0,088	0,089	0,074	0,083	4
AEC3D 5-5	0,071	0,071	0,076	0,083	0,081	0,075	9

Source: Original research results

Table 11 details the formation of the Gaussian factor and notes that the average demand for rows was the parameter that had the greatest weight in the final score of the clusters followed by the amount of SKUs and factor sku_caixa.

Table 11: Gaussian factor formation – Final selection

Parameter	skus	linhas_média	sku_caixa	fator_linha	linha_caixa
Average	0,083	0,083	0,083	0,083	0,083
Standard deviation	0,020	0,031	0,019	0,007	0,013
Gaussian factor	0,245	0,376	0,227	0,088	0,158

Factor G. Norm.	0,224	0,344	0,207	0,080	0,145

Source: Original research results

In comparison with the univariate analysis, the Clustering technique associated with PCA and Autoencoder was able to group the SKUs considering all variables leading to the formation of better balanced clusters as shown in table 12. The clusters formed maximized the number of SKUs and the average demand for lines without proportionally increasing the number of boxes and the maximum demand for lines. In the univariate analysis, the increase in the average of lines resulted in a proportional and, in some cases, even higher increase in the number of boxes and maximum lines.

Table 12: Comparison of univariate and multivariate analysis

group	skus		Boxes		linhas_média		linhas_max	
Total demand	12.204	(100%)	44.055	(100%)	21.348	(100%)	46.468	(100%)
PCA2D 1-3	8.232	(67%)	24.880	(56%)	13.028	(61%)	24.045	(52%)
PCA3D 1-3	8.625	(71%)	25.593	(58%)	13.253	(62%)	24.443	(53%)
AEC3D 3-3	9.791	(80%)	27.297	(62%)	11.650	(55%)	18.130	(39%)
abc 5	2.440	(20%)	28.858	(66%)	16.221	(76%)	40.519	(87%)
xyz 1	2.441	(20%)	11.492	(26%)	5.805	(27%)	11.108	(24%)
Whistle 5	2.440	(20%)	9.754	(22%)	7.588	(36%)	19.595	(42%)
Stock 5	2.440	(20%)	29.872	(68%)	15.600	(73%)	39.785	(86%)

Source: Original research results

From the point of view of intralogistic projects, the multivariate analysis was efficient in forming clusters of SKUs with more appropriate characteristics for the proposed system: automatic system AS/RS shuttle type for storage of boxes and separation of fractional units in a GTP station. Considering the same need for boxes and maximum lines, the clusters of the multivariate analysis were able to meet a higher number of SKUs and a higher average demand of lines when compared to the clusters of the univariate analysis. In other words, for the same level of investment in inventory and handling equipment, the groups formed through the Clustering technique can meet a higher demand. It is then considered that the research objective was achieved once unsupervised machine learning techniques were identified that can assist in the problem of grouping SKUs and it was verified that the groups formed through these techniques have more adequate characteristics than those formed by univariate analysis.

Final Considerations

The problem of grouping SKUs was examined through the application of conventional techniques based on univariate analysis such as ABC, COI and XYZ. Unsupervised machine learning techniques such as PCA, Clustering and Autoencoder were also applied. These techniques proved to be more efficient in the grouping of SKUs since they consider all the variables of the base such as demand, demand variation and dimensional characteristics of the products. In the selection of the best groups, the Gaussian AHP method was used, which proved effective in identifying the most appropriate clusters for the proposed automation system. The present research proved the potential application of unsupervised machine learning techniques in the development of intralogistics automation projects and initiated some studies that can be deepened in future research.

The following limitations of the study have been identified and they represent opportunities for future work:

- application of these techniques in more databases and in databases in other sectors of industry,
- investigation of the techniques used, but considering parameters different from those adopted in this research,
- the investigation of other techniques that can also assist in the solution of the proposed problem,
- the division of the base into training and testing to prove that the grouping and assignment of SKUs remains adequate over time,
- the application of supervised machine learning techniques for classifying new SKUs and rearranging existing ones according to the dynamic conditions of a warehouse operation.

References

Bahrami, Behnam & Piri, Hemen & Aghezzaf, El-Houssaine. (2019). Class-based Storage Location Assignment: An Overview of the Literature. In Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2019): 390-397.

Boehmke, B., & Greenwell, B.M. (2019). Hands-On Machine Learning with R (1st ed.). Chapman and Hall/CRC. Disponível em: https://doi.org/10.1201/9780367816377. Acesso em 20 set. 2022.

Chen, Lu & Langevin, André & Riopel, Diane. (2007). The storage location assignment and interleaving problem in an automated storage/retrieval system with shared storage. International Journal of Production Research 48: 991–1011.

De Koster, René & Le Duc, Tho & Roodbergen, Kees Jan. (2007). Design and Control of Warehouse Order Picking: A Literature Review. European Journal of Operational Research 182: 481 – 501.

E. H. Frazelle. (2002). World-Class Warehousing and Material Handling. McGraw Hill, New York, New York, USA.

Favero, L. P. L., & Belfiore, P. P. (2017). Manual of data analysis: statistics and multivariate modeling with excel, SPSS and stata. Elsevier, Rio de Janeiro, Rio de Janeiro, Brazil.

Fernandes, J. & Campilho, Raul & Pinto, Gustavo & Baptista, A.. (2019). Intralogistics and industry 4.0: designing a novel shuttle with picking system. Procedia Manufacturing 38. 1801-1832.

Fontana, Marcele & Leyva Lopez, Juan & Cavalcante, Virgínio & Solano, Jaime. (2020). Multi-criteria assignment model to solve the storage location assignment problem. Investigacion Operacional. 41. 1019-1029.

Fontana, M. E., López, J. C. L., Cavalcante, C. A. V., & Noriega, J. J. S. (2020). Multi-Criteria Assignment Model to Solve the Storage Location Assignment Problem. Revista Investigation Operational 41(7): 1019 – 1029.

Gu, Jinxiang & Goetschalckx, Marc & Mcginnis, Leon. (2007). Research on warehouse operation: A comprehensive review. European Journal of Operational Research 177: 1-21.

Gu, Jinxiang & Goetschalckx, Marc & Mcginnis, Leon. (2010). Research on warehouse design and performance evaluation: A comprehensive review. European Journal of Operational Research 203: 539-549.

Hausman, Warren & Schwarz, Leroy & Graves, Stephen. (1976). Optimal Storage Assignment in Automatic Warehousing Systems. Management Science 22: 629-638.

Kofler, Monika. (2015). Optimising the storage location assignment problem under dynamic conditions. Tese de Doutorado. Universidade de Linz, Linz, Austria.

Lorenc, Augustyn & Kuźnar, Małgorzata & Lerher, Tone. (2021). Solving product allocation problem (PAP) by using ANN and clustering. FME Transactions 49: 206-213.

Mirzaei, Masoud & Zaerpour, Nima & De Koster, René. (2021). The impact of integrated cluster-based storage allocation on parts-to-picker warehouse performance. Transportation Research Part E: Logistics and Transportation Review. 146. 102207.

Reyes, J.; Solano-Charris, E.; Montoya-Torres, J. (2019). The storage location assignment problem: A literature review. Int. J. Ind. Eng.Comput. 10: 199–224.

Rouwenhorst, B. & Reuter, Boris & Stockrahm, V. & Houtum, Geert-Jan & Mantel, R.J. & Zijm, W.H.M.. (2000). Warehouse design and control: Framework and literature review. European Journal of Operational Research 122: 515-533.

Santos, M. dos; de Araújo Costa, I.P.; Gomes, C.F.S. (2021). Multicriteria decision-making in the selection of warships: A new approach to the AHP method. Int. J. Anal. Hierarchy Process 13: 147–169.

Silva, Allyson & Coelho, Leandro & Darvish, Maryam & Renaud, Jacques. (2020). Integrating storage location and order picking problems in warehouse planning. Transportation Research Part E Logistics and Transportation Review: 140.

Stojanović, Milan & Regodić, Dušan. (2017). The Significance of the Integrated Multicriteria ABC-XYZ Method for the Inventory Management Process. Acta Polytechnica Hungarica 14: 29-48.

Zrnic, Nenad & Popović, Tamara & Milojević, Goran & Kosanić, Nenad. (2021). A Survey of Research on Industry 4.0 in Intralogistics. X International Conference “Heavy Machinery-HM 2021”, Vrnjačka Banja, 23 – 25.