POLITECNICO DI TORINO

Collegio di Ingegneria Gestionale e della Produzione

Corso di Laurea Magistrale in Management Engineering

Application of machine learning clustering techniques for performance analysis in ocean freight transport: a Ferrero case study

Supervisor

Prof.ssa Claudia Caballini

Candidate

Lorenzo Vinciguerra

A.A 2024/2025

Contents

Chapter 1 – Logistic & supply chain management

Chapter 2 - The distribution system in freight transport

Abstract

This thesis applies machine learning clustering techniques to the analysis of ocean freight transport flows, with the aim of identifying patterns and inefficiencies. The method is applied to the international logistics network of the Ferrero Group. The primary objective of this research is to identify inefficiencies within a portion of the supply chain, i.e. the distribution process, by examining three crucial performance indicators: the total storage days products spend in warehouses, the overall lead times from production to shipment, and the associated risk of aging stock that might lead to expiration. To this aim, two clustering methods, K-Means and DBSCAN, were applied, in order to group products and highlight similar logistics behavior. The missing values were handled via median imputation, numerical variables were standardized and categorical attributes were label-encoded. Then K-Means and DBSCAN were executed: K-Means chose k = 2 (silhouette 0.504), yielding a dominant baseline cluster (~95%) and a compact hotspot (~5%) with markedly elevated storage days (~660), longer lead times (~32) and high ageing exposure (~1,123). DBSCAN was a density-based check and validated the hotspot, revealing dense local pockets and outliers associated with lanes, temperature classes or pack types. In a simple case, reducing the average storage in the hotspot from about 660 days to 330 takes the network-wide average back by 12 percent, hence where a particular tactic (e.g., beginning a new booking earlier) is more likely to pay the greatest payoff. The work demonstrates that this critical point is composed of the smallest percentage of SKUs products whose excessive time in the warehouse and high exposure to shelf-life expiration risk disproportionately drive the overall system's inefficiency. This led to the key conclusion that, for this specific small segment of products, focusing on targeted improvement strategies, such as revising inventory policies or optimizing shipping schedules, can yield significant gains for the company's logistics and supply chain performance.

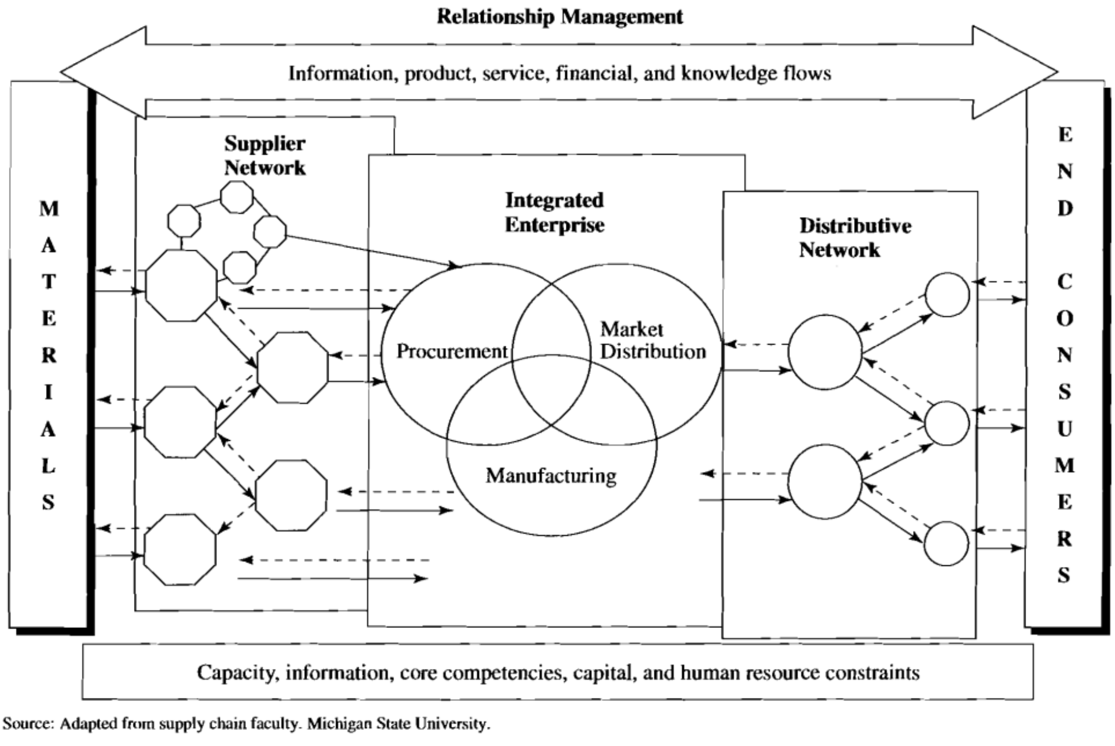
Chapter 1 - Logistics and the supply chain

1.1 Definition

The roots of logistics are to be found in the military sphere. Careful study of the movement of troops and supplies can become the key factor for victory or defeat in a conflict; the victories of Napoleon are an example. His different way of managing the supply chain gave his troops greater mobility than his opponents. Even in the Second World War and more recent conflicts, logistics played a leading role (Brandimarte & Zotteri, 2004). Once the Second World War ended, the knowledge and techniques were transferred from the military sphere to the context of industrial management. [1]

From here was born the so-called industrial logistics, which can be defined as the set of activities within a company that deal with managing the flow of information and materials starting from the procurement of raw materials and up to the production phases of the product, packaging, transport, distribution, storage, and after-sales service.

Logistics has three major operational areas: procurement, support for production, and product or service distribution to the final market. [2]



Source: Adapted from supply chain faculty. Michigan State University.

Figure 1 - Generalized supply chain model, R. Bowersox, J. Closs and M. B. Cooper

Procurement is the one that purchase raw materials, semi-finished goods, and final products. In the production support phase, the main objective is to schedule production through a master production plan that considers both the product demand and the availability of raw materials. Product distribution focuses on the journey of the finished product from the production plant to customers; therefore, it is important to highlight that the most significant relevance is the availability of the product in the times, places, and quantities required by the consumer (Bowersox, Closs, and Cooper).

Logistics can also be defined as the process that merges the different activities of the supply chain into a single integrated operation. It usually represents one of the most significant costs for the company, but if logistics activities are carried out efficiently, they can constitute a success factor over competitors. Having a good logistics strategy means having low costs and meeting the requirements and level of service requested by the customer (Bowersox, Closs, and Cooper).

1.2 Supply chain

The supply chain can be defined as "a set of three or more entities (organizations or individuals) who are directly involved in the upstream and downstream flows of finances, products, services, and information from a source to a customer" (Defining supply chain management, Mentzer, 2001). [3]

Industrial logistics was born to solve problems encountered within the enterprise; it developed over time, optimizing and integrating the various processes of the internal supply chain or internal supply chain. Nowadays, with the advent of the Net Economy, this traditional internal supply chain approach within company walls can be said to be superseded to make way for the extended supply chain, that is, a chain where not only the internal processes of the enterprise are integrated, but all the processes that involve the company in question and external subjects to it. The entire logistics network is considered a unit with the extended supply chain. The most common problems in this case concern all interface activities that link the different enterprises and often generate inefficiency. Another aspect to bring to light concerns the different ways in which companies compete with each other. In the past, individual companies competed with one another, but with the advent of the Net Economy, the different logistics networks compete. In this new scenario, the company inserted in the most efficient, structured, and integrated logistics network will not necessarily be the best company, but it will have the upper hand.

Therefore, the supply chain is characterized by a complex network created around all the organizations participating in a particular process. To create a better idea about the supply chain is to consider it a group of companies, structures, and entities that make it possible to bring the product or service to the consumption market, coordinating and synchronizing all the activities to manage resources at the best. Moreover, as Stevens (1989) shared, supply chain management has the task of finding the best compromise between a good level of service and a low unit cost. [4]

In Figure 2, a general scheme is shown of all the factors that must be present to ensure efficient and effective supply chain management, starting from the involvement of the various functions and the external environment through bonds of trust, integration, risk sharing, and cooperation to obtain value for the end customer and competitive advantage for the company.

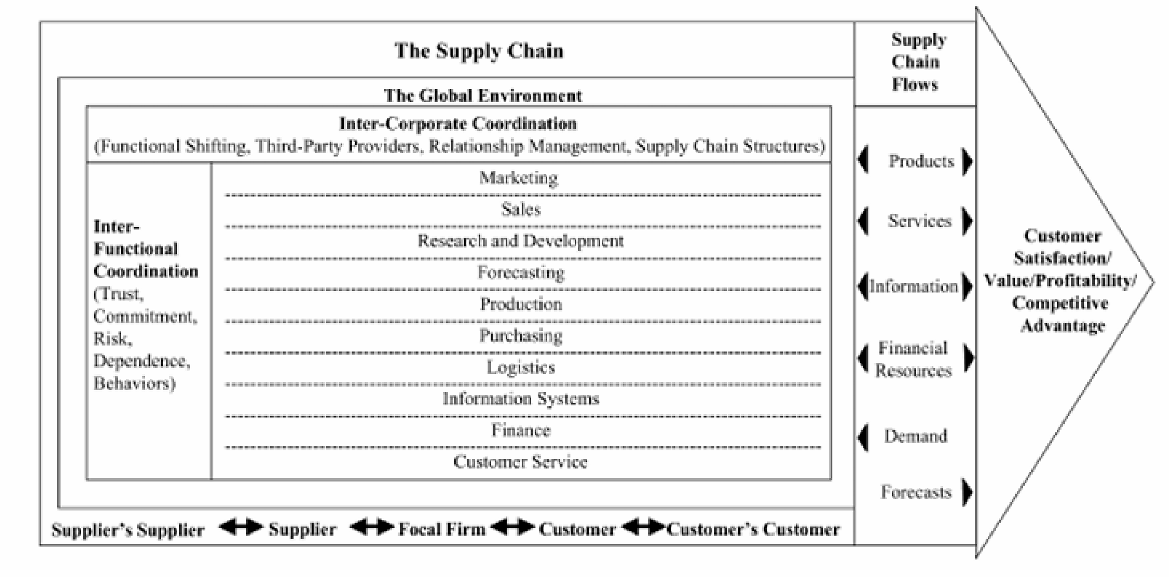


Figure 2 – A model of supply chain management, Bowersox, Closs, and Cooper.

1.3 Integrated logistics

Successful supply chain management involves sharing information, risks, and benefits. Sharing information facilitates internal and external processes within the organization under examination; for this reason, it is important to try to integrate as much as possible the company functions and those concerning external suppliers and customers.

All these integration problems find a solution in integrated logistics, which aims to optimize the trade-off between efficiency and effectiveness, that is, to minimize the total cost of logistics activities as a whole, given a service level objective to be guaranteed .

Another related factor that determines the proper functioning of a company's supply chain, and consequently the success of integrated logistics, is the collaboration with suppliers and third parties. If two organizations trust each other over time, they can plan innovative projects around common aims more easily.

Integrated logistics can be summarized as cooperation that begins with joint planning and ends with joint control activities to evaluate the performance of the members of the supply chain and the entire supply chain system.

This concept of integrated logistics must be shared by all the parties that constitute the supply chain; the common objectives must be the same and must be made known and shared, starting from the production of the product up to its distribution. It is also important to integrate the third-party logistics provider where present (Mentzer, 2001). Where a third-party operator is present, outsourcing of the management of logistics activities occurs. In some cases, companies prefer that specialized distribution and storage companies manage these functions on their behalf. The advantages lie that these third-party companies have specialist skills in the sector and therefore can obtain more favorable rates and better service, leaving the company in question more time and resources to focus on its core business.

The main processes that should be included in this context of integrated logistics are: procurement, production, order fulfillment, customer service, demand forecasting, product development and its commercialization.

The strategies to be followed to obtain a competitive advantage can be cost leadership or differentiation. Cost leadership involves following a strategy that reduces costs to the greatest extent possible, thereby lowering prices for final consumers compared to competitors. The second strategy is determined by creating something different, special, and unique. One must try to differentiate one's product or service as much as possible from that of competitors. Although opposed, both can create a competitive advantage and lead the company to success.

In the integrated logistics there are many objectives and the importance of each one depends on the context in which the company is considered. The first objective is responsiveness, understood as the company's ability to meet customers' needs quickly. The second is reducing variance. The latter is present in every logistical operation; for example, damaged goods are a source of variance, a transport that, taking the wrong route, arrives at the wrong place or late, or an unforeseen interruption of production, etc. A solution to mitigate variance is safety stock, or, regarding variance in deliveries, using faster transport in the event of unexpected delays; these practices have a very high cost, reducing them to the indispensable minimum is good. The third objective to be achieved is to reduce inventory. Stock in warehouses represents a massive cost for the company, but often a high inventory can hide underlying inefficiencies that could be eliminated by adopting integrated principles. In Logistics the objective is to reduce inventory to a minimum while continuing to meet the desired level of performance. The fourth objective is the consolidation of shipments where transport costs represent extremely high logistics costs, so this objective is of primary importance. The unit transport cost decreases as the quantity moved increases and with distance. Therefore, problems arise where small and timely deliveries are required. Therefore, an innovative system and multifaceted coordination are necessary to consolidate deliveries. The fifth objective consists of continuous quality improvement, since a product is delivered defective, it is customary to make a return or exchange. Every time a product has a quality that is not acceptable to the customer, all the costs of direct logistics are added to the costs of reverse logistics. It is therefore extremely. Important to reduce return flows to a minimum by increasing product quality. The sixth objective is support in the product life cycle, it consists of ensuring assistance starting from the delivery of the product to the moment of its destruction and disposal. It is therefore essential to structure and plan. The reverse logistics phase is effective and efficient in a world where recycling is now of primary importance; it is essential to provide a method for the disposal of goods sold (Bowersox, Closs, and Cooper).

1.4 Logistics flows

During logistics activities, different types of flows pass between the various actors. The supply chain includes materials, information, services, money, and technical knowledge.

Material flow: It is the main flow of logistics activities and may be made up of finished products, components, semi-finished goods, and raw materials that must be used for production, transport, stored, and finally distributed to the market. Starting from purchasing raw materials and supply chain activities, generating added value by moving stock. Where and when necessary (Bowersox, Closs, and Cooper). Also belonging to this type of flow is reverse logistics. All return, disposal, and waste recycling practices are part of reverse logistics; they could be a solution.

There are many environmental problems. Companies are aware of the growing attention paid to pollution and climate issues; consequently, reverse logistics is gaining increasing importance. Services such as assistance, maintenance, product repair, and returning defective items are now indispensable requirements for customers.

Therefore, reverse logistics is also fundamental with regard to customer service: in today's context of growing and aggressive competition, a company that does not provide this type of service is decidedly disadvantaged compared to competitors.

Information flow: It is made up of all data exchanges between the actors of the supply chain. This flow is fundamental to ensuring logistics activities' coordination, efficiency, and effectiveness. Where proper sharing of information is present, a better overall performance of the supply chain is ensured; moreover, there is a reduction in uncertainty and less Misalignment of operational activities (Bowersox, Closs, and Cooper). If the decision-maker could benefit from perfect information, then their choices would be universally optimal. Nevertheless, it is impossible to enjoy perfect information since there will always be a margin of error due to human intervention (Brandimarte & Zotteri, 2004). To facilitate the exchange of information, it is now essential that the logistics network have an efficient information system, capable of providing the desired answers in the shortest possible time, returning a consistent and reliable output at any time.

Technical knowledge flow: it is essential since it is through this flow that the company's know-how is transferred. This means that those who possess relevant information are often reluctant to share it. When the employee who guards the key process information retires or changes jobs, if their knowledge has not been transmitted to another member internal to the company, this information is lost; this problem is known in the literature as knowledge loss. When this situation occurs, the solution must be sought mainly in corporate culture and organization: for example, in a company where the mindset is to establish a family atmosphere and mutual trust with colleagues, employees will be more willing to share their knowledge.

Financial flow: it is fundamental to sustain the company's business. Profit is the ultimate objective that guides supply chain management and the company's logistics strategy. However, financial performance is measured using indicators that are different from those. Used for logistics performance, this can lead to misalignment and inefficiencies. Therefore, the finance function must be in maximum harmony with logistics activities to prevent this kind of misalignment from causing problems in operations.

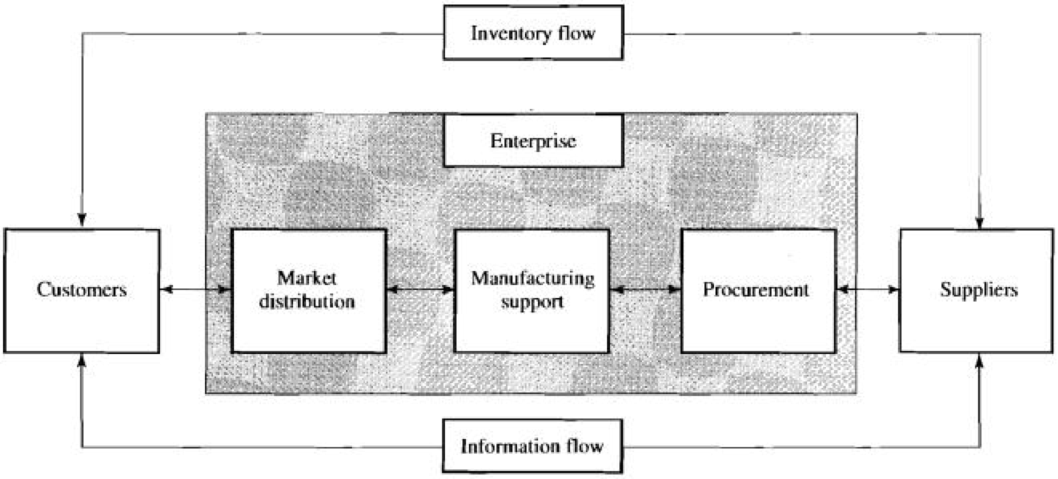


Figure 3 - The flows in the supply chain. Source: Bowersox, Closs, and Cooper.

1.5 The structure of the supply chain

As explained earlier, an integrated supply chain can generate added value for the final customer; in this case, all functions involved in logistics activities are synchronized, and in this way, there is a saving in costs and many benefits in terms of both efficiency and effectiveness. This paragraph will briefly introduce the main functions that must work cooperatively and coordinate to perform the supply chain activities. The leading logistics operations can be grouped into the following areas:

- order fulfillment

- inventory management

- transport

- warehousing, materials handling, and packaging

- network of facilities.

Order fulfillment includes all the activities necessary to process customer orders; this is a critical phase and influences the success of the subsequent logistics operations. During order processing, the company directly contacts the customer, who expresses their needs, requirements, and any customizations through the order. Once the order has been received, it is very important that all the information obtained be shared with the other functions; this can be very easy if the function that manages orders has appropriate IT support and a secure and reliable information system. One of the main performance indicators of this function is order fulfillment time; in fact, if time is saved in the order intake phase, one can opt for a slower transport without the risk of failing to deliver by the set date. Furthermore, alongside the time saving, there could also be a cost saving, since faster transports correspond to higher rates while slower transports correspond to more modest rates (Bowersox, Closs, and Cooper).

About the management of warehouse inventory, the main objective is to meet customer requests; product availability must be ensured, given a certain level of service, and with the lowest possible stock level. Keeping a large quantity of goods, to have the product available at any time, can entail an excessively high cost and prove inefficient under other aspects, such as product obsolescence. Here too, having the correct information is extremely important. If it is not possible to store the entire product requirement, the best thing to do is to give priority to the most profitable products.

Another important factor that a company must evaluate in inventory management is the profitability of its customers.

“The profitability of a customer's business depends upon the products purchased, volume, price, value-added services required, and supplemental activities necessary to develop and maintain an ongoing relationship.” In this sentence by Bowersox, Closs, and Cooper, the factors determining a customer’s profitability are listed. Since a company’s economy is based on its main customers, it makes sense to think of an inventory management strategy that adapts to the needs of these customers (Bowersox, Closs, and Cooper).

Furthermore, since inventory derives from production or supports it, these two elements cannot be managed separately and must always be coordinated. Economic factors must also be considered when making decisions relating to inventory management. In fact, warehouse stock generates many costs, including, first and foremost warehousing costs related to personnel, structural costs, maintaining particular conditions, etc.; following are the costs entailed by the obsolescence or rapid perishability of goods; and finally, the opportunity cost of capital investment, which therefore cannot be employed in other ways (Brandimarte & Zotteri, 2004).

Another important phase of integrated logistics consists of transport from the production plant to the warehouse or from the warehouse to the final customer. Transport represents the highest cost of distribution, usually equal to 30–60% of the total costs of the latter. A company can carry out transport activity in different ways; for example, it can sign several contracts with different suppliers or a single contract and thus establish a unique relationship of trust with the supplier. To choose the best option, three different aspects must be considered: cost, delivery speed and consistency.

"The transport cost is the payment for shipment between two geographical locations and the expenses related to maintaining in-transit inventory" (Bowersox, Closs, and Cooper). It is essential to choose the transport solution that minimizes the cost of the entire logistics system and not only the transport cost. In other words, it is not a rule that the cheapest way to transport goods implies the lowest total cost.

The second factor to consider is the speed with which it is necessary to deliver the goods. The type of good transported also becomes important to evaluate the importance of making a delivery in more or less tight times. For example, for foodstuffs with short shelf life or for which there is a minimum shelf life, it may be essential to make delivery as soon as possible. On the contrary, transport speed assumes less importance if the good under analysis is non-perishable and there are no other needs for which delivery must be made in the shortest possible time. What must be done before choosing one solution over another is a careful comparison between the customer's needs and requests and the costs in question. There is a trade-off between cost and delivery speed, both essential factors. The objective is to find the right balance between the two to optimize transport effectiveness.

The third aspect to consider is delivery consistency, as defined by Bowersox, Closs, and Cooper : "Consistency of transportation refers to variations in time required to perform a specific movement over several shipments." If the same trip always has the same duration, it becomes extremely easier to carry out scheduling and forecasting activities; on the contrary, if the duration of the trip varies from time to time, it is difficult to have accurate planning, and this generates a series of problems that reverberate throughout the supply chain. To conclude, it can be stated that transport speed and delivery consistency determine transport quality. Finding the right compromise between cost, speed, and delivery consistency is the main objective in choosing the transport system (Bowersox, Closs, and Cooper).

The functions of material handling, warehousing, and packaging are essential for reducing the costs of the logistics process. By material handling, we mean all the activities connected to the movement and storage of goods within the warehouse. The machinery used for handling impacts both efficiency and management costs. Moreover, material handling is fundamental to facilitate transport and accommodate the customer's needs. If the movement of goods is easier, operators will carry out handling activities faster, and consequently, there will be savings in terms of time and cost. The expenses for moving goods are minimized when the goods are moved and then arranged in a different solution the least number of times. For unloading and loading goods, too, material handling is critical since if the goods are arranged effectively, the times and difficulty of the operation will be reduced (Bowersox, Closs and Cooper). Warehouses store goods that must be available when the customer requests them.Warehouse management makes decisions regarding the location, number, layout of distribution centers, receiving methods, and methods of preserving goods. If warehousing activities are carried out effectively, there will be an economic benefit in reducing overall logistics costs. An important choice that every company must make is whether to own facilities and thus keep the warehousing function in-house or rely on a third-party operator expert in the sector. Both options have advantages and disadvantages: entrusting all storage activity to a third party would certainly prove more efficient than keeping the warehousing function in-house; however, on the other hand, there could be serious integration problems between the two parties, which will have to find a common interface. Therefore, in making the choice that best fits the specific situation, it will be advisable to consider the aspect related to cost and the added value that the two solutions generate (Bowersox, Closs, and Cooper).

Packaging has the function of identifying, containing, and protecting the product and facilitating its distribution. For consumer products, packaging can also be a choice dictated by marketing, since identifying the good precisely from its packaging is an aspect not to be underestimated. The packaging must protect the product from many factors: humidity, heat, pressure, oxidation, animal infestations, molds, or insects. Packaging must be robust enough to protect the good at every moment of distribution. In addition, to facilitate handling, products are grouped into increasingly larger units; there are usually three levels of packaging: the first, called primary packaging, contains the product; the second groups together several packages in the case where the primary packages are small in size; and the third groups together in a single unit several primary or secondary packages. Once the third level of packaging is reached, a unit load is obtained, composed of multiple packages inside it. This aggregated unit facilitates product handling and further protects the individual packages, which will unlikely be damaged thanks to the greater resistance conferred.

Packaging is a necessary cost that must be borne to confer greater efficiency in product distribution. In conclusion, when the functions of warehousing, material handling, and packaging work in an integrated and cooperative way with the other logistics functions, the result is an increase in the speed and ease with which the product flow advances, starting from production to distribution, and a consequent reduction in costs (Bowersox, Closs, and Cooper).

Decisions regarding the network of facilities are crucial for the performance of the supply chain for three reasons: they involve a significant capital investment, a medium/long/term commitment, and have a significant impact on the efficiency and costs of logistics operations. The need for network planning may concern the design of new facilities and the redesign of existing ones.

The network's facilities are mostly warehouses, production plants, transit points, cross-dock operations, and points of sale. The reasons why it is often convenient to redesign the structure of the existing network include inefficient operations (e.g., high costs, bottlenecks), the introduction of new products or services, changes in environmental requirements, changes in the volume handled, etc. Network design must be redesigned continuously, adapting it to the new needs that arise over time. Poor network design can negatively affect system performance. Network design requires careful study of the geographical characteristics of the territory. Only through the continuous redesign of the network can one have an innovative supply chain worthy of a company that wants to occupy a leadership position in the sector (Bowersox, Closs, and Cooper).

1.6 The actors of the supply chain

In this paragraph, the different actors of the supply chain are introduced. These actors, who may be physically located in places distant from each other, receive and ship physical flows through the transport system while exchanging information flows electronically.

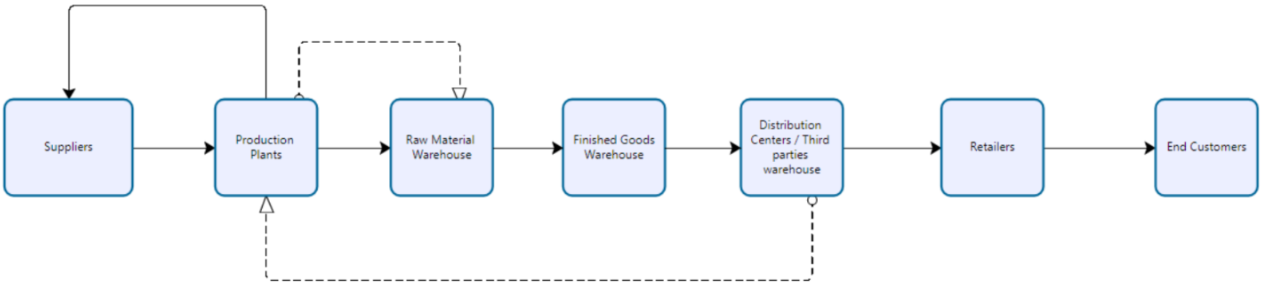


Figure 4 – Supply chain Actors, graph generated with Bizagi

Warehouses, as already explained above, store goods while awaiting their transport. They have a fundamental function in reducing the impact of variation in demand and supply along the supply chain. They can be cataloged according to the goods they contain; thus, according to this classification, warehouses can be distinguished as:

- finished product: if the output of production, ready to be delivered to the point of sale where the final customer can purchase it, is stored;

- raw materials: if inside, there are the materials necessary to begin production, subassemblies, or components that have not yet entered the production process;

- semi-finished: if products are contained that have already undergone a processing but are waiting to continue their path through the production process;

- maintenance: if all tools, machinery, or supplies used in production are stored, that do not become part of the finished product.

It should be noted that an item's classification may vary according to the level of the supply chain at which it is located; for example, flour will be considered a finished product for the supplier but a raw material for the confectionery company.

The role of production plants is to process the quantity of goods necessary to meet market demand under the constraint of production capacity. In factories, raw material enters as input and the finished product exits as output; therefore, during the production process, added value is generated. Even in production, it is necessary to identify logistics requirements and the necessary conditions to coordinate this function as well as possible with the other activities of the supply chain. It has been shown that the greater the level of coordination and cooperation between procurement and production, the lower the costs incurred. Secondly, joint production and inventory management design also yields huge benefits (Bowersox, Closs, and Cooper).

Suppliers procure the raw materials, semi-finished goods, components, or parts that the company in question needs for the production of the finished product. Market competitiveness encourages suppliers and firms to establish long-term relationships based on trust and cooperation; it becomes increasingly important to integrate suppliers within the logistics network to maximize efficiency. It follows that the choice of suppliers is critical and must align with the strategic orientation. If such a choice is successful, value will be generated in terms of higher profits, greater competitiveness, and competitive advantage for both the company and its suppliers. The concept of integration with suppliers is expressed exhaustively by the following statement: “To develop a successful partnership and to reach the mutual goals between the partners, firms must have business communications associated with the positive atmosphere of discussions, interdependence, and shared constructive expectations.

Depending on the type of product in question, the company and its logistics chain will position themselves in a given market. Based on the product under analysis, it will be the company’s task to satisfy the needs of the market’s consumers in order to reap the most tremendous possible success. The customers in question can be of many types depending on the kind of company under analysis and its level within the supply chain; for example, the customer can be another company, a wholesaler, a distribution center, or a retail store.

Companies often deal with a varied quantity of products intended for different markets and yet use the same strategy in the supply chain. This fact generates a problem since, if the product's characteristics do not fit the logistics strategy adopted, optimal cost and service cannot be achieved.

Another important aspect to consider is the service offered to the customer that can be considered in different ways; for example, inventory management is understood as product availability when needed. There is no better or worse measure of customer service; all the indicators used can find strengths and weaknesses. For example, an indicator of the level of customer service offered could be the number of orders processed within the determined times or the number of deliveries to the customer made by the set day. Indeed, having a higher level of stock helps to guarantee a higher level of customer service, since protection from uncertainty is greater. The costs that the company may have to bear if inventory is insufficient to guarantee the level of service offered are mainly three: back-order costs, lost sales, and lost customers. These costs are generated when a stockout occurs and are very difficult to evaluate. In conclusion, the decision on what level of service to guarantee depends on corporate choices and especially on the type of market in question; in fact, if stockout costs are estimated to be very high, it is advisable to guarantee a higher level of service.

1.7 Green logistics

Nowadays, attention to the environment has become a very important and felt topic both by companies and consumers. Many companies have realized the negative impacts that their activity can generate on the environment that surrounds them and, consequently, have begun to worry about minimizing damage. Logistics stands out among the functions involved in this movement toward ecological approaches, hence the expression “green logistics.”

Green Supply Chain Management is defined by Shang, Lu, and Li (2010) as follows: “Green Supply Chain Management involves financial flow, logistics flow, information flow, integration, relationships, and environmental management, promoting efficiency and synergy among partners, facilitating environmental performance, minimal waste, and cost savings. Therefore, it is an important source for the competitive advantages of organizations.” The main difference between conventional supply chain management and green supply chain management is about the fact that the former deals mainly with optimizing the aspect linked to economy, while the latter, in addition to economy, also seeks to optimize environmental and social aspects, as shown in Figure 5 below. [5]

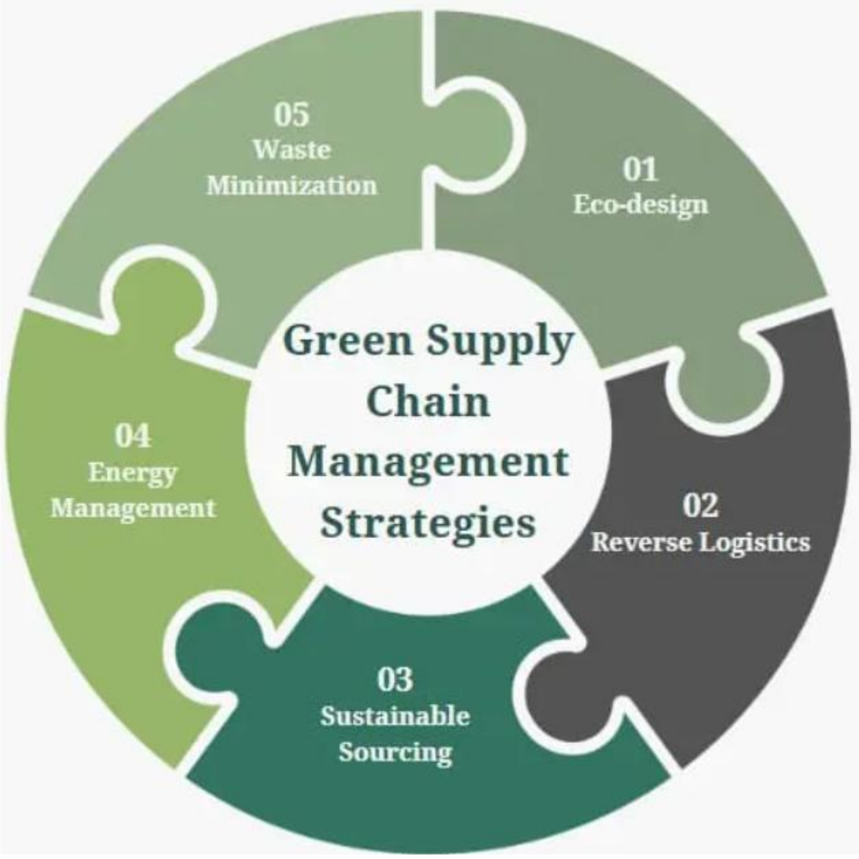


Figure 5 - The principles of green supply chain management, Green Supply Chain Management Strategies [6]

- Eco-design: developing products, in a way that minimizes their impact by keeping an eye on energy usage, material intensity and how simply the parts can be repurposed or recycled once the product reaches the end of its life.
- Reverse logistics: manage the journey of products from customers back, to the producer enabling reuse, remanufacturing, recycling or responsible disposal.
- Sustainable sourcing: Choose suppliers that meet social standards and give preference to inputs that are renewable, biodegradable or that demand fewer resources.
- Energy management: reduce the chain's energy consumption by deploying technologies that optimize transport and warehousing and shifting a larger share of the energy mix toward renewables.
- Waste minimization: Reduce waste, by streamlining production, reclaiming and recycling materials and opting for packaging.

It can be said that the primary objective of any company is profit; therefore, ways must be found to ensure that eco-sustainable solutions generate gains for the company; in this way, corporate interests would be aligned with environmental ones. A first factor very important for the development of green logistics is represented by the regulations in force in the country, since it has been demonstrated that where incentives and taxation on the ecological nature of a company's operations are present, firms develop more rapidly solutions more compatible with the environment (Virendra Balon, 2019). Recovery of investments is also a strategy that is favorable both to the company and to the environment, since where there are investments that are now unused, it is convenient to look for a way to still derive some benefit from them. For example, in the case of excess inventory or inventory now unusable for normal commercialization, it may be advantageous to sell it at a lower price rather than destroy it. Or, if a company owns a machine now out of use, before scrapping it, it is worth checking that it cannot in fact be converted in any other way. Recovery of investments consists mostly in the recovery and reuse of materials and thus includes reverse. Another factor that works in favor of green logistics is the fact that consumers themselves have become very sensitive to respecting the environment; moreover, they are also willing to pay higher prices to have an eco-sustainable product. Therefore, it is convenient for companies to respect the environment so that their reputation does not negatively affect turnover (Virendra Balon, 2019). Green design is also a solution to the problem, since if one starts to design the product by reconciling the consumer's needs with respect for the environment, great success will be obtained; a prime example is the hybrid car that was developed when the price of oil had risen, to meet customers' needs (Virendra Balon, 2019). [7]

In concrete terms, some aspects on which green supply chain management works are:

- eco-sustainable vehicles that minimize diesel consumption and polluting emissions. In addition to ecological means, it is also important to train drivers so as to avoid driving behaviors that negatively affect consumption.

- Intermodality, that is, replacing, when possible, road transport which is the most used with other more ecological modes of transport.

- Recyclable packaging and reducing its size.

- Using "green" warehouses that use renewable energies and are built with eco-sustainable materials.

- Using software, programs, and tools that make it possible to optimize trips to reduce unnecessary movements to a minimum, trying not to have vehicles travel empty.

- Collection, recycling, and disposal of waste.

In conclusion, it can be said that the benefits of green logistics are not only linked to the environment. Although high initial investments are often necessary, the company has a return in terms of optimization of resource management, a smaller quantity of waste, an increase in efficiency and productivity, a reduction in storage and transport costs, an increase in economic profitability, and an improvement in corporate reputation.

1.8 The market in food logistics

The food market has always played an important role in the European economy, especially for Italy.

The Italian food industry grew by 10% in 2023 [8], reaching a total value of €90 billion, confirming the positive trend over the last decade. From 2014–2024, Italy's food industry has grown far more rapidly than the GDP.

Over this period, the Italian food industry's turnover increased from about €53 billion in 2012 to over €90 billion in 2023, a cumulative rise of around 70% [9]. For Italy foreign trade has also proved very important, with which the majority of the firms analyzed achieve 50% of their turnover. Therefore, it can be deduced that it is precisely internationalization that makes the Italian food industry a profitable sector for the country's economy. In this scenario and in relation to the continuously growing trend of the population that is recorded every year, together with the need to serve the consumer while maintaining the same standards, the role of food logistics becomes extremely relevant.

Food products must be shipped all over the world facing different climates and weather conditions that subject them to physical and environmental stresses. Therefore, a good management of logistics processes, including storage, packaging and transport is paramount.

Despite the food supply chain being intrinsically similar to the automotive one in terms of complexity and capillarity, the logistics solutions developed for the second sector's products were not applicable to the first sector's products. In fact, the great difference between these two sectors is dictated by the perishability of the products in question, which, in the case of the food industry, therefore requires specific requirements. [10]

In addition, in the food industry, quality and traceability requirements are also important; supporting this statement, it has been shown that the relationship with cost for perishable products is less important than for non-perishable products [11]

Natural conditions also dictate a series of consequences and aspects that must not be neglected during procurement and production. Because of all these aspects, in the food supply chain, the relationships between the different actors strictly depend on one another and therefore integration and cooperation are increasingly important issues in this context. It can be stated that the characteristics of food products create a unique scenario for the different actors of the supply chain with the need for specific requirements for logistics activities [12]

1.9 The products in the food industry

A classification that can be made to distinguish the different products deriving from the food industry is the one relating to the temperature at which they must be stored. This characteristic plays a prominent role in the operational choices that must be made during the phases of storage, transport and preservation at the point of sale [13].

Depending on the temperature needed to transpiort the products, three different supply chains can be distinguished: refrigerated, frozen and ambient temperature.

Refrigerated products: they must be kept at a temperature slightly above the freezing point, therefore between 0 °C and 10 °C. This process slows down, but does not stop, the spoilage of foods. This means that refrigerated products can be stored for short periods of a few days or at most two weeks.

Frozen products: frozen products are products subjected to a process where foods are brought to temperatures between -7 °C and -12 °C and then stored at temperatures between -10 °C and -30 °C. Once foods are defrosted there is a loss of nutritional and organoleptic values. An alternative method where freezing takes place very quickly and effectively is the deep-freezing process; in this case foods are brought in a very short time to -18 °C, thus creating micro-crystals of water that do not damage the biological structure of the products, preserving in this case their nutritional values and flavors. This method makes it possible to significantly extend the preservation of foods.

Ambient-temperature products: this includes the entire category of products that do not require low temperatures for their preservation; this category includes all dry products, among which also chocolate, honey, coffee, and almost all fruit and vegetables. The distinction of products into these three sets is extremely important because, if a certain temperature is necessary for the preservation of the products, it must be guaranteed during all the supply chain processes. [14]

1.10 The actors in the food supply chain

The supply chain in the food market is composed mostly of four types of actors: primary producers, industrial producers, salers and retailers.

Primary producers: those who supply the raw materials which in this case derive from agriculture and livestock. In this first phase of the supply chain, the role of traceability and the quality of foods is extremely important. In particular, primary producers should collaborate with the other actors of the supply chain to guarantee continuous production and a good level of service.

Industrial producers: they include all the plants where the raw materials are processed through countless processes until reaching the finished product. Here, added value is created through the transformation of raw materials, the packaging, and the packaging of the finished product ready to be marketed. In this phase, due to the perishability of the product, monitoring the speed of delivery and the geographical position of the other actors in the chain is very important.

Salers: distributors who store and transport products between primary and industrial producers. For these actors in the chain, transport costs and the study regarding demand forecasting prove to be very important. A problem that wholesalers must face is the combination of different products with different preservation and shelf-life characteristics.

Retailers: those who deal with selling to the final consumer. In this phase of the chain all the problems regarding distribution and the availability of the product on the shelf must be managed. The most important thing within the food supply chain is that the different actors interact with each other with a network approach, since the quality of the finished product depends on how it is handled during all the processes of the chain. In addition, it is important to underline that the needs of retail can influence the behaviors of the entire supply chain [14]

1.11 Transport in the food industry

The food industry can use different transport options. The most common is road transport, widely used since the food industry requires temperature control and the maintenance of the level of quality of products, which is easy to preserve with this type of transport. Sea transport is also commonly used; however, it is more suitable for large quantities of goods and for products with long preservation. Rail and air transport are less used. Warehouses also play a fundamental role in the food industry: to serve small markets with various types of products it is indeed very important to create consolidation clusters. In this way the vehicles will travel with higher saturation, obtaining environmental and economic advantages. An important factor for the development of logistics clusters is also the position of the warehouse, which must be located at a strategic point in order to serve the entire market effectively [15]

Chapter 2 - The distribution system in the supply chain

2.1 Transport in the supply chain

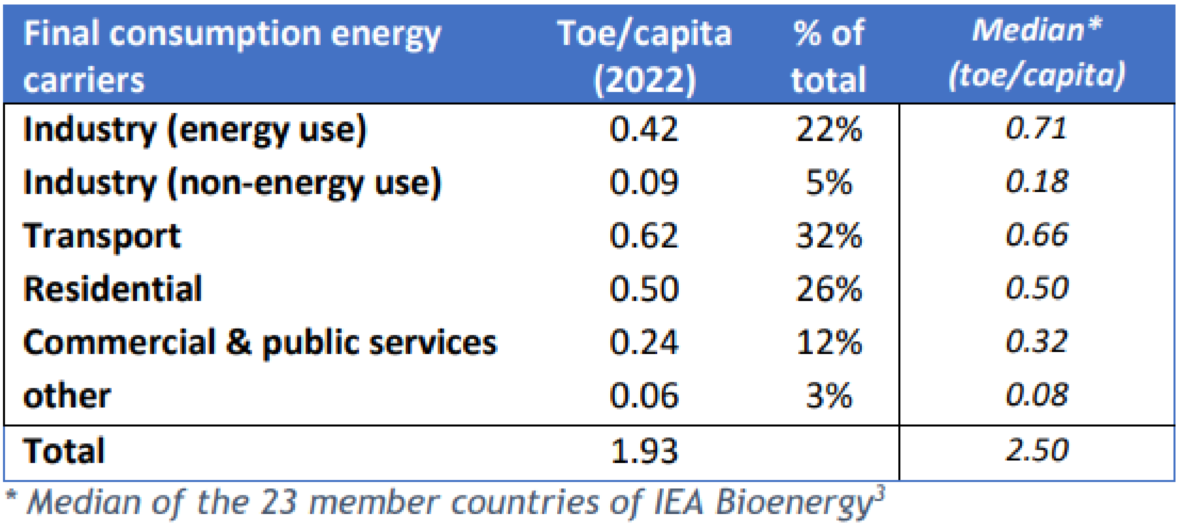
Before the deregulation of transport in 1980, there was minimal difference between different carriers; in fact, service and prices were generally the same.

After 1980, when price flexibility was introduced, the services offered became many and varied. Nowadays, each company offers a different service; this has been made possible not only by free price choice but also by the development of technology, which provides unique and innovative solutions to support transport activity. Thanks to information technology, today it is possible to track transport in real time, determine its position, and estimate delivery time, aspects that have led transport to acquire greater importance and no longer limit itself only to moving goods from one point to another. Without reliable transport, logistics activities would be put to a severe test (Bowersox, Closs and Cooper).

Transport consumes time, financial, and environmental resources. First, during the transport phase, goods are inaccessible; the quantity of product transported is called transit inventory. The goal of logistics managers is to reduce to a minimum the in-transit inventory, which is not accessible until delivery has been made. Second, transport requires financial resources to cover costs generated by the driver's work, fuel consumption, vehicle use, infrastructure use, investments, and necessary management. Finally, transport uses environmental resources both directly and indirectly. Transport can be attributed as one of the largest consumptions of fuel. Although means of transport are increasingly efficient, the total consumption of fossil fuels in the sector remains high.

In the Table below is reported the final energy consumption per capita in Italy by end-use sector (2022) and comparison with the median of the 23 IEA Bioenergy member countries (Refers to the 23 countries that participate in the IEA Bioenergy Technology Collaboration Programme; the "Median (toe/capita)" is calculated across these 23 countries, not all IEA/OECD members).

Transport is the largest contributor (0.62 toe/capita; 32%), followed by residential (0.50; 26%) and industry—energy uses only (0.42; 22%). Italy's total final energy use amounts to 1.93 toe/capita, about 23% below the IEA Bioenergy median (2.50 toe/capita). The gap is widest in industry (0.42 vs 0.71) and in commercial & public services (0.24 vs 0.32), while residential matches the median (0.50) and transport is close to it (0.62 vs 0.66). [16]



Final consumption energy carriers	Toe/capita (2022)	% of total	*Median** *(toe/capita)*
Industry (energy use)	0.42	22%	*0.71*
Industry (non-energy use)	0.09	5%	*0.18*
Transport	0.62	32%	*0.66*
Residential	0.50	26%	*0.50*
Commercial & public services	0.24	12%	*0.32*
other	0.06	3%	*0.08*
Total	1.93		*2.50*

** Median of the 23 member countries of IEA Bioenergy*[3]

*[**toe** =tonne of oil equivalent a standard unit that measures energy by comparing it to the energy content of one metric tonne of oil.]*

Table 1 - Distribution of the final consumption of energy carriers by sector in Italy. (Source: IEA, World Energy Balances and Renewables Information, 2024; GSE.)

Figure 6 shows how Italy's transport energy mix evolved from 2010 to 2022, highlighting diesel's dominance, the smaller shares of gasoline, LPG and natural gas/biomethane, and the temporary Covid-related drop in 2020.

Italy's transport energy is still led by diesel (about 65%). Gasoline has fallen to about 23%, while LPG and natural gas (including biomethane) are smaller shares (about 5% and 3%). In 2020 there was a sharp drop (~19%) in total fuel use because of Covid; the fuel mix then returned close to earlier levels. For freight, this means road haulage depends mostly on diesel, so changes in diesel use track goods movement. Practical steps for the near term are better logistics efficiency and fleet renewal, using higher biodiesel/HVO blends where engines allow, more biomethane for truck fleets, and shifting some long-distance flows to rail or short sea. [16]

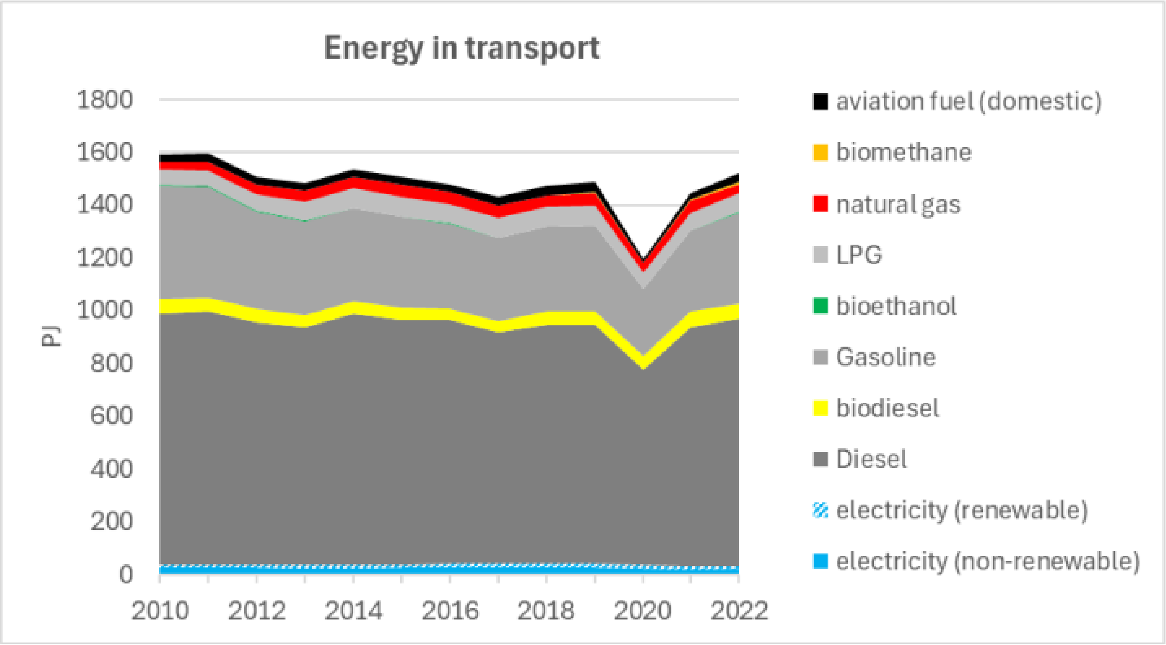


Figure 6 - Transport energy mix in Italy (2010–2022): shares of diesel, gasoline, LPG, and natural gas/biomethane. (Source: IEA, World Energy Balances and Renewables Information, 2024; GSE.)

The renewable share in transport stays around 4–5% (with a temporary spike in 2020). Biodiesel provides most of it; ethanol is small and has declined; biomethane is growing (about one-fifth of transport gas in 2022). Electricity is still a small share (~2.2%) of transport energy and is used mainly in rail; road electric vehicles are rising but were ~0.14% of transport energy in 2022. For freight, this suggests near-term cuts in emissions come mainly from advanced biofuels/HVO and biomethane for heavy trucks, plus more rail/intermodal on electrified routes; battery or RFNBO options will matter more as technology and infrastructure expand.

In transport, lower running costs are a key benefit. Electric vehicles are cheaper to run than petrol or diesel cars because electricity per km is lower and maintenance is simpler. For heavy road freight, advanced biodiesel/HVO and biomethane can be produced from local waste streams, improving energy security and helping fleets hedge against oil price swings. [17]

Figure 7 below shows how much of Italy's transport energy already comes from renewables and from electricity (2010–2022), so we can see where savings are happening today and where therewhere there is room to grow.

The renewable share in transport stays around 4–5% (with a temporary spike in 2020). Biodiesel provides most of it; ethanol is small and has declined; biomethane is growing (about one-fifth of transport gas in 2022). Electricity is still a small share (~2.2%) of transport energy and is used mainly in rail; road EVs are rising but were ~0.14% of transport energy in 2022. For freight, this suggests near-term cuts in emissions come mainly from advanced biofuels/HVO and biomethane for heavy trucks, plus more rail/intermodal on electrified routes; battery or RFNBO options will matter more as technology and infrastructure expand. (Data: IEA, World Energy Balances and Renewables Information, 2024; biomethane from GSE.)

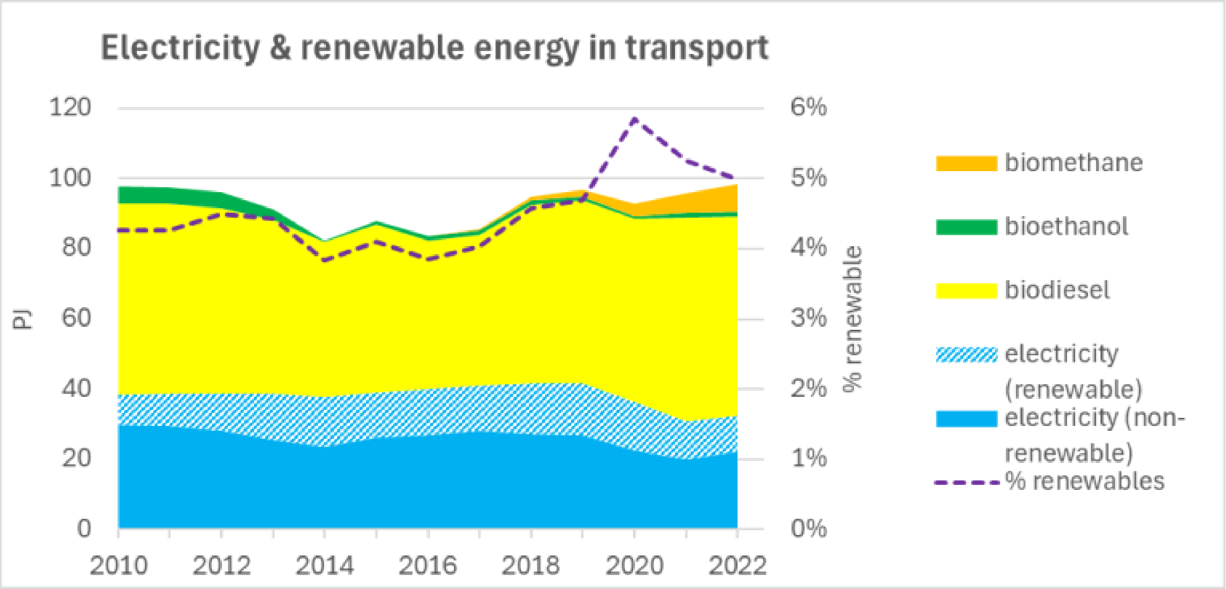


Figure 7 - Electricity and renewables in Italian transport (2010–2022): renewable share of transport energy and use of electricity (mainly rail). (Source: IEA, World Energy Balances and Renewables Information, 2024; GSE.)

Indirectly, transport contributes to atmospheric and noise pollution through combustion. However, from the chart in Figure 8, it can be seen that starting from 2005, there has been a progressive decrease in greenhouse gas emissions. [18]

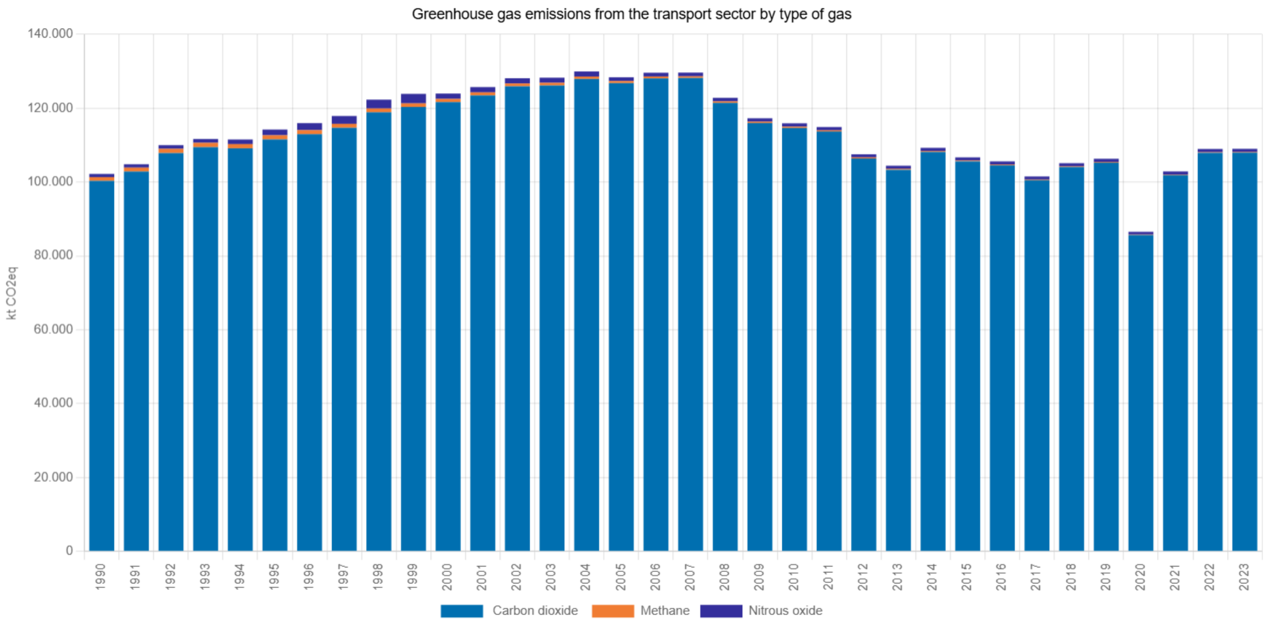


Figure 8 - Greenhouse gas emissions from the transport sector by type of gas and share of transport in the total (Source: ISPRA - National Greenhouse Gas Emissions Inventory).

The data in the Figure 8 are taken from the National Greenhouse Gas Emissions Inventory reported under the UNFCCC framework. The figures refer to total national emissions net of removals from the LULUCF sector (Land Use, Land-Use Change and Forestry). The historical series has been recalculated from 1990 based on methodological updates adopted in the IPCC sectoral estimation approach.

As regards transport efficiency, two aspects must be considered, namely economies of scale and economies of distance. Economies of scale in transport consist in reducing the cost per unit of load as the quantity being carried increases; that is, if a container's capacity is fully used during a shipment, the cost per cubic meter will be lower than in a situation where capacity is used only halfway. It can also be observed that vehicles able to carry higher volumes have a higher cost than those that can carry smaller quantities. For example, rail transport will have a lower cost than air or road transport. This occurs because the costs generated by transport are mostly fixed and therefore do not change based on variations in the volume carried. A similar reasoning can be made for economies of distance. In this case the cost per unit of load decreases as the distance traveled increases. Here too, if the distance is greater, fixed costs are spread over more kilometers and therefore the cost per kilometer will be lower. The objective of transport is thus to maximize volume and distance while respecting the guaranteed customer service (Bowersox, Closs and Cooper).

2.2 Transport modalities

There are five types of transport that can be used, with very different characteristics from one another: rail, road, water, air, and pipeline. The market share of each type of transport can be assessed by considering system mileage, the volume carried, and the category of goods carried. A standard unit used in the field of transport is the ton per unit of distance (Bowersox, Closs and Cooper).

Rail transport

In the past rail transport played a leading role in the sector. The feature that dictated the success of this delivery mode is the economy of large shipments; service frequency is also considered a strength. Up to the Second World War railways were the main means of transport. In the postwar period, to the detriment of rail, the number of roads began to grow, which gave a great boost to road haulage. The peculiarity of rail transport is its high fixed costs due to equipment investments and track access rights. On the other hand, variable costs are very low. With the advent of road transport, the goods carried by rail have been limited to a few specific categories; for the most part,raw materials of extraction companies and products with low value density are carried, where the cost per ton-mile becomes critical. Moreover, rail is often used in intermodal systems, combining its cost advantage over long distances with the flexibility of road for the first and last mile.

Road transport

Road transport is the most common and flexible transport mode. The ability to reach almost any point of origin and destination without the need for trans-shipment makes it particularly advantageous for short and medium distances and for shipments requiring capillary distribution. The fixed costs of the road system are relatively low compared to rail and water, while variable costs are higher due to fuel, labor and maintenance. The main strengths are frequency, accessibility and regularity of service; the main weakness compared to rail and water is cost over long distances and exposure to congestion and weather.

Water transport

This type of transport is the oldest. A distinction can be made between deep sea transport and transport on inland navigable waters. Transport on inland waters has maintained a constant share of volume in recent years. The size of the network, which is the most limited among the transport modes, has also remained unchanged and is expected to remain so in the near future. The strength of boat transport is the ability to carry extremely large volumes.

In terms of fixed costs, water transport is a middle way between rail and road. An advantage for companies that operate these carriers is that access rights are subsidized by government. The main disadvantage, however, is the low speed of transport. Also, the fact that it is almost always necessary to add a land leg to connect points of origin and destination is not a factor in favor of maritime transport. Nevertheless, this transport proves the best choice when one intends to carry a large quantity of goods at low cost and there are no stringent time constraints for delivery.

Air transport

Although air transport offers a series of extremely advantageous strengths, it remains the least used mainly because of its high costs. Even though this type of transport generates lower fixed costs than rail and water, it entails considerable variable costs caused by fuel consumption, user fees, high labor intensity and maintenance. In addition, the main limitations of air transport are the limited capacity of the vehicles, the availability of aircraft and the payload threshold. On the other hand, the main advantage of this mode is the speed of delivery; for example, a coast-to-coast shipment can be performed in just a few hours by air, whereas it would take days with the other types of transport. Firms choose this mode essentially when they need to provide a certain level of service that is able to justify the high cost. At present there is no single category of goods to which air transport is destined; this mode is preferred for high-value, time-sensitive or perishable cargo, urgent spare parts, and in general for flows where lead time is a critical competitive variable.

Pipeline transport

Pipeline transport is used to move fluids (oil and derivatives, natural gas) and, in certain cases, slurries. Its main advantages are continuity of flow, very high frequency and reliability, and minimal exposure to congestion and weather. Its limitations are the restricted range of products that can be moved and the rigidity of routes once the infrastructure has been built.

Comparing the different modes on the basis of operating characteristics, as regards reliability(understood as the ability to meet guaranteed service levels and times) the safest mode is pipeline, thanks to the fact that it is not affected by weather conditions and does not suffer from congestion problems, followed by road transportWhen considering capacity, such as the size of the load, water transport is the best option. And finally, as for frequency, that is, the number of trips that can be made in a unit of time pipelines are in first place because of their continuous flow, immediately followed by road (Bowersox, Closs and

Cooper, 2002). As for fixed costs, the most expensive transport is pipeline, followed by rail and water. Road and air enjoy low fixed costs. With regard to variable costs, air is the most expensive, followed by road. Rail, water and pipeline prove economical from the standpoint of variable costs. Maritime transport makes up the largest share of EU freight, among 5 transport modes: maritime, road, rail, inland waterway and air during the last decade.

Figure 9 shows that the share of maritime freight transport reached its lowest point of the decade in 2023, at 67.4% (4 823 billion tonne-km), after falling for 4 consecutive years. Most notably, there were decreases by 0.7 percentage points (pp) between 2020 and 2021 and 0.6 percentange points between 2019 and 2020. Maritime freight transport recorded its highest share in the last decade in 2015, with 69.6%. [19]

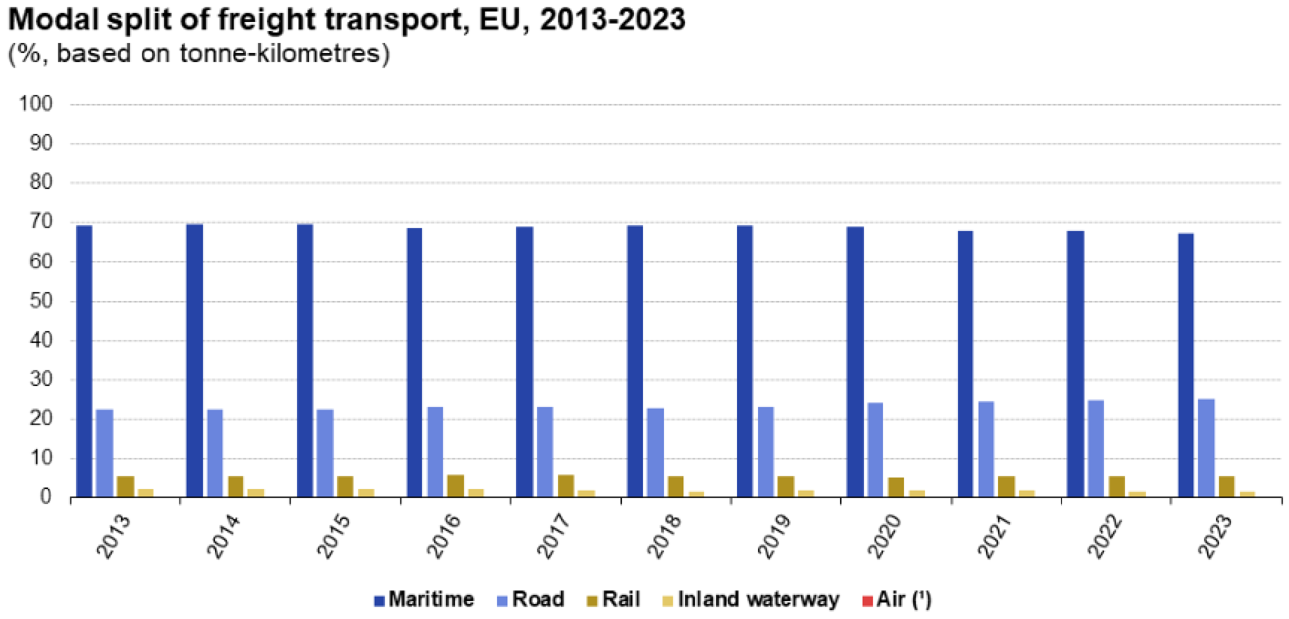


Figure 9 - Modal split of freight transport in the EU (Source: Euostatistics - Freight transport statistics)

The share of road transport in the total EU freight transport performance reached a peak of 25.3% (1 807 billion tonne-km) in 2023, after an increase of 0.4 pp compared with 2022. Over the period 2013-2023, the share of road transport had its lowest point in 2014, at 22.4%.

The share of rail in the freight transport performance was relatively stable over the period 2013-2023. A peak was observed in 2016, at 6.0%, while a low point was reached in 2020, at 5.2%. In 2023, the share was stable compared with 2022 (5.5% or 391 billion tonne-km).

The share of inland waterway in total freight transport performance was also relatively stable over the period 2013-2023. A peak was observed in 2013-2014, at 2.2%, while the lowest level of 1.7% was reached in 2018. In 2019, the share slightly increased to 1.8% and remained at this level until 2021. In 2022 and 2023, a low point was attained at 1.6% (116 billion tonne-km).

The share of air in freight transport performance remained at 0.2% (14 billion tonne-km in 2023) during the whole period from 2013-2023. (Euostatistics - Freight transport statistics, Modal split of freight transport in the EU).

At a global level, the shares of transport modes are different because they reflect the geographical conditions in which transport systems operate, such as country size, coastline, distances, and population density. We can highlight their behavior through Figure 10 below, which considers five countries with different environmental characteristics.

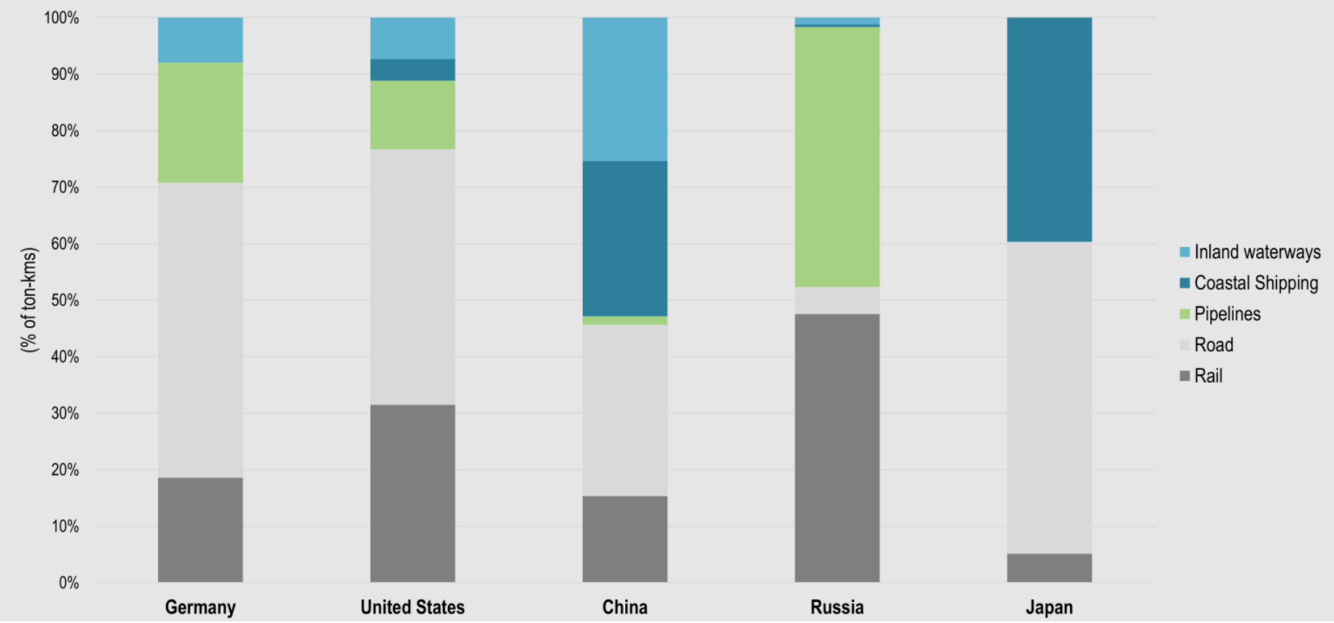


Figure 10 - The Geography of Transport Systems, Modal Share of Freight Transportation, Selected Countries [20]

The share of each mode of transport depends on geography, so on how big a country is and how much coastline it has. In Europe, China, and Japan, road and coastal shipping are most of the ton-km, while rail is the main mode in the United States and Russia. A country's resource base also affects the cargo mix: a high supply of natural resources leads to more rail than manufacturing and services. In the United States and Russia, all the industrial activities are spread across areas with different roles like manufacturing, farming, and the generation of resources, which creates long-distance rail flows. There is also strong use of pipelines to move oil and natural gas, especially in Russia, and also in the United States. In Europe and Japan, high population density and shorter distances favor trucks.

Japan has almost no inland waterways because of its geography. Coastal transport is very important in Europe, China, and Japan because many people live near the coast and some regions are separated by seas (the Baltic and Mediterranean in Europe, the Yellow Sea in China; Japan is a group of islands). A long coastline usually means more coastal shipping, but in the United States the efficient rail and road systems give coastal shipping a smaller role. In Russia, the long Arctic and Pacific coasts are thinly populated, and parts are closed by ice in winter. (Modal Share of Freight Transportation, Selected Countries) [20]

2.3 Warehouses

"A warehouse is generally viewed as a place in which to keep or store inventory." This is the definition of warehouse provided by Bowersox, Closs and Cooper (2002). Warehouses can be distinguished as factory, regional and local warehouses. They mainly perform two service functions of a warehouse according to Arnold, Chapman and Clive (2007) [21]:

1. Store and protect the goods for a long period until they will be needed; these warehouses are called general warehouses and, for example, can store goods with seasonal sales.

2. Make the goods available for distribution. These warehouses are called distribution warehouses. They are very dynamic; in fact, there is a very high frequency of goods in and out. The goods arrive in large uniform lots, are stored for a limited time, and then reorganized into small heterogeneous lots to be shipped to different destinations.

There are several reasons that justify the storage of goods; the main reasons are the following:

- economy of scale in transport: larger volumes entail lower transport rates and therefore it is convenient to collect at the warehouse goods destined for the same area to consolidate shipments and thus obtain a saving

- balancing supply and demand: since often the inflow of goods and their outflow toward the market do not coincide, a warehouse is needed to compensate for this discrepancy and balance flows

- support for production: storage of raw materials and semi-finished goods allows production to continue smoothly even in the presence of variability in supplies

- postponement: delaying the final configuration or packaging of a product until closer to the demand point allows for greater flexibility in serving customers' specific requests

- safety stock: keeping an additional quantity of inventory to protect against uncertainty in demand and lead times

- service level: having product availability near the customer allows to improve delivery times and the perceived level of service

The costs associated with warehouses are many and can be grouped into the following categories:

- fixed costs: costs for buildings, depreciation of structures, insurance, taxe

- variable costs: personnel for handling, energy, equipment operation

- inventory holding costs: costs related to capital tied up in stock, obsolescence, shrinkage, and spoilage

- handling costs: costs generated by the internal movement of goods, receiving, picking, and shipping

- information system costs: hardware, software, and maintenance of the WMS and other systems necessary to manage the warehouse.

The design of a warehouse must consider the layout (arrangement of areas and aisles), the storage systems (shelves, racks, automated warehouses), the equipment for handling (forklifts, conveyors), and the logic for slotting and picking. The objective is to minimize the movements necessary to receive, store, and ship goods, while guaranteeing the required service level. Among the main performance indicators are:

- average orders processed per hour

- average lines picked per hour
- picking accuracy (%)
- average inventory level and inventory turnover
- average dwell time by SKU
- space utilization (%)

The operations that are commonly performed within a warehouse are receiving, management of internal movements, shipping, and storage. Receiving includes the set of activities necessary to unload goods arriving from suppliers or production plants, check quantities and condition, register entry in the information system, and prepare the product for storage or immediate cross-docking. It is important that the receiving area have to be sized according to the expected inbound flow and equipped with the necessary equipment to unload safely and quickly.

Internal movements concern the moving of products from one portion of the warehouse to another: from receiving to storage locations, from storage to picking, from picking to consolidation and shipping. Shipping includes order consolidation, packing, labeling, and loading onto the outbound vehicle. It adds up to things such as transport lead times, carrier schedules, and even customer delivery windows. Storage is the stage where the product stays in its assigned area until a request for picking has been made. The storage system (pallet rack, drive-in, gravity flow, etc.) chosen is determined by product rotation, quantity per SKU (stock keeping unit), and space available.

2.4 Costs

The total logistics cost related to transport and storage is the result of a compromise among several elements. On the transport side, the main cost items are the line–haul cost (proportional to distance and volume), the cost of pickup and delivery, the terminal handling cost, and any costs related to invoicing and collection. On the storage side, the main items are fixed structure costs and variable costs linked to inventory and handling. The optimal solution is obtained by minimizing the sum of these costs with respect to the network decisions (number and location of warehouses), the consolidation policy, and the chosen transport mode.

From an economic point of view, a warehouse exists when the overall cost of a system with storage is lower than without storage, given the same service level. This includes: the benefits of economies of scale in transport, reduction of stockouts thanks to risk pooling, and the possibility of postponement. Storage also introduces costs and risks (obsolescence, capital tied up), so the decision must be supported by quantitative analysis.

2.5 Different types of inventory

Inventory can be classified according to its reason for existence and its function in the flow. Below are described the main categories:

- Cycle stock (or lot-size inventory): this is the portion of inventory that results from purchasing or producing in batches larger than the immediate demand. It cycles down as consumption proceeds and is replenished when the next order or production lot arrives. The cycle stock level depends on the economic trade-off between ordering/setup cost and holding cost.

- Safety stock (or buffer): This is the extra number saved to mitigate uncertainty in demand and/or supply lead time. It does so in an attempt to prevent stockout at a given level of service. Safety stock determinants consist of the demand variability, lead-time variability, average lead-time and the desired fill rate. In normal-approximation, safety stock is expressed as a product of the standard deviation of demand during lead time.

- Seasonal inventory: Pre-planned in advance of predictable peaks in demand or event planning (e.g., promotions, seasonality, plant shutdowns). The capacity utilization is smoothened and inventory risk minimized for times or seasons with high supply risk (which in turn results in higher holding cost) by manufacturing or buying ahead of time.

- Pipeline (in-transit) inventory: Things having left the shipping point and are moving toward the receiving point. It cannot be consumed until it arrives and is equal to the average demand times transit time. Stock in pipelines is larger at longer distances and to slower modes of transport.

- Decoupling stock: stock being held between two stages of a process to dissipate fluctuations and enable some degree of separation between upstream and downstream operations. It safeguards the system that gets interrupted fairly little and avoids the requirement for perfectly synchronized control.

- Speculative inventory: inventory that one buys in advance to gain advantageous price increases, quantity discounts or to hedge against supply disruptions. Although it might be beneficial, it raises the chances of obsolescence and further commitment of capital.

In practice, there could be different uses for the same physical units (e.g., on-hand stock will have cycle stock and safety stock). The above categories are useful for management and accounting purposes, when justifying the inventory maintained, as well when selecting policies (orders, reorders, reviews).

2.6 Ocean freight transports

Sea freight has developed through three main stages. The first, up to about 1850, depended on sailing ships whose speed depended on wind; ships were similar in build and carried bulk and liquid goods in many holds, and offloading in unequipped ports could take days. The second stage, after 1850, was led by steamships, because of industrial growth and the need for faster turnover; bigger ships and new sea jobs appeared as global trade grew fast. The third stage began in the early 1900s with the cable telegraph and then wireless telegraph, which let ships communicate with ports and with each other while at sea. In 1906 the ocean liner Republic sent the first message to shore by telegraph. [22]

Figure 11 - A container ship is docked, and its cranes are loading containers on board at the quay. [23]

Since ancient times the sea has been an important source of food and a route for the transport of people and goods by means of many types of boats. Different civilizations developed, improved, and specialized the ships used, and centuries of sea history brought big progress in the capacity, speed, and safety of ships. The year 1967 marks the start of the use of containers; the business began with a transpacific service, but soon spread around the world, helping to speed up economic growth. In 2007, on the 40th anniversary of the event, twenty-three of the top companies in sea transport set up the Container Shipment Information Service (CSIS), an organization that aims to provide a single source of data and information that shows the advantages of containerization in world economic growth. Goods are transported worldwide mainly by specialist shipping companies that operate through a head office and a network of local agents. The goods carried can be divided into five main categories that decide the type of ship and the special port facilities to be used: liquids, gas, dry bulk cargo, general cargo, and container cargo. In the latter case the shipping companies have not only their own ships but also the containers. The various types of goods can also be put in different land transport units, such as trucks and rail wagons, which are then carried on ferries or ships.

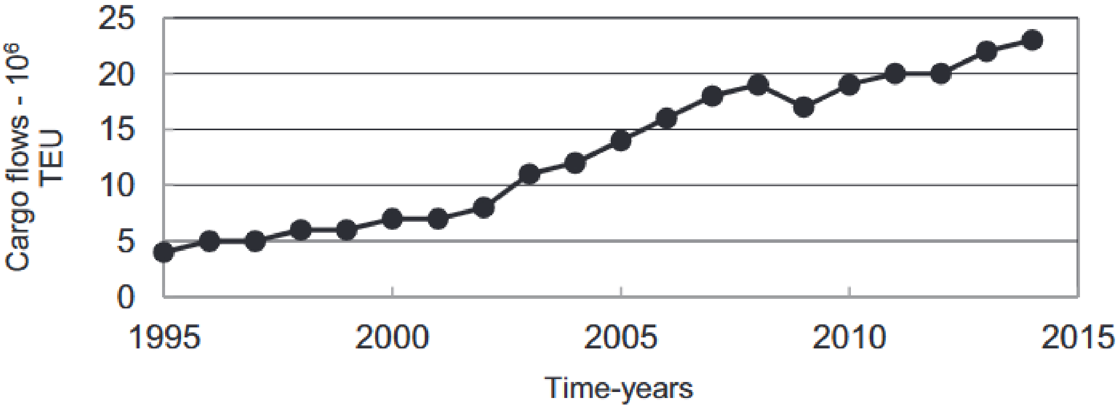


Figure 12 - Development of containerized cargo flows over time (Period: 1995-2014), Bart Wiegmans [24]

Freight transport services provided by deep-sea shipping have been steadily growing over time, as shown in Figure X. To give a sense of scale, let us now look at a few key numbers that illustrate how much this part of the supply chain has expanded. [24]

In 1956, loading cost about $5.86 per ton, with containers it dropped by more than 90% to about $0.16. Container terminals were present in roughly 1% of countries in 1966 and in about 90% by 1983.
Ship loading speed rose from about 1.3 tons per hour before containers to about 30 tons per hour after their adoption. Today, around 90% of commercial goods move in containers. There are more than 17 million containers in circulation, completing over 200 million trips per year, and the active fleet includes over 6,000 container ships. These numbers explain why containerization became the standard for global trade. [25]

2.7 Containers

The need to increase the volume and speed of freight transport and to reduce management costs and risks led to the use of standard, smaller containers that can be loaded onto the three main means of transport: truck, rail car, and ship. This system began to operate around 1930 and was widely used during the Second World War for military needs.

The container is defined as a freight transport unit, equipped with devices (corner fittings) that allow lifting and fastening on the relevant means of transport. The standard container sizes are 20 feet and 40 feet and have the following characteristics:

- 20': width 2.438 m – height 2.591 m – length 6.058 m – internal volume 33 m^3
- 40': width 2.438 m – height 2.591 m – length 12.192 m – internal volume 66 m^3

Containers are made of steel for strength; moreover, they must not be too heavy in relation to the load carried. ISO (International Organization for Standardization) standards also classify three other longer containers used mainly in the United States and in certain geographic areas:

- 45' (length 13.7 m), introduced in 1980
- 48' (length 14.6 m)
- 53' (length 16.2 m), introduced in 2007

Special trade needs have led to the development of other types of containers (tank containers for the transport of liquids, refrigerated containers for goods that need to travel at controlled temperature, etc.).

The 20-foot container is used as the basic unit of measure for the calculation of slots on container ships and bears the acronym TEU (Twenty-foot Equivalent Unit).

Standardization also extends to container identification, which uses an alphanumeric code of four letters (prefix) and seven numbers (serial number), the last of which is used for data checking. Code markings follow special rules that depend on the company to which the containers belong and identify the owner, nationality, type, and size of the containers.

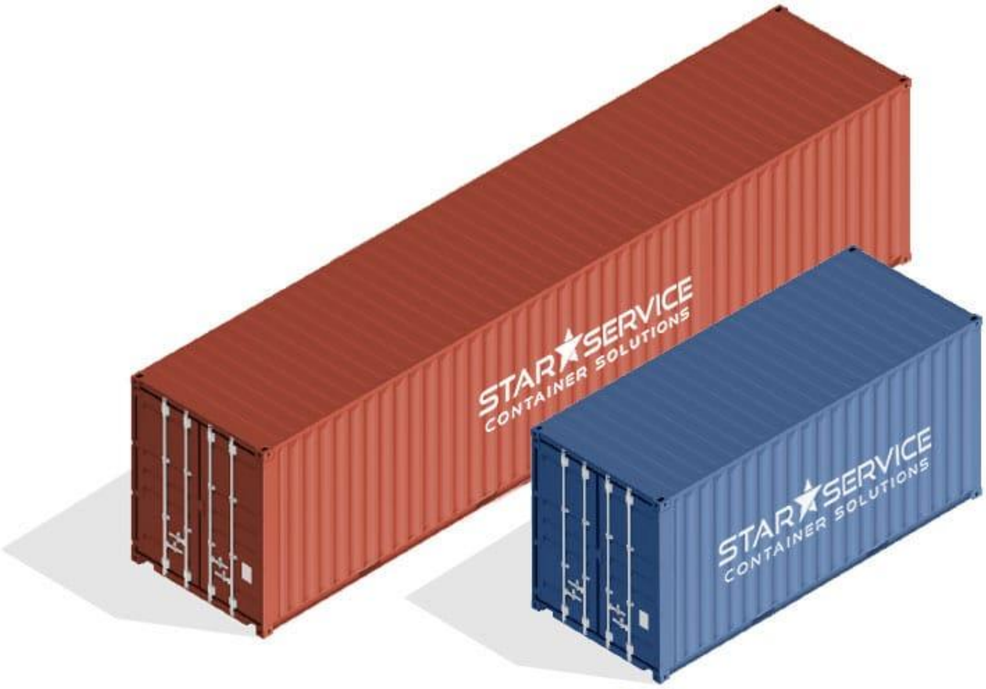


Figure 13 – A 20-foot and a 40-foot container, Star Service SRL

2.8 Terminal container overview

A container terminal is a fixed facility where maritime containers are managed and handled in order to change their mode of transport. Generally, a container terminal is associated with the handling of containers between containerships and between ships and land vehicles, such as trains or trucks; in these cases, it is defined as a maritime container terminal.

Similarly, the handling can take place between land vehicles, typically between a train and a truck; in that circumstance, it is defined as an inland container terminal, usually an interport. Since terminals are the hub of container transport, they play a special role in national and international transport routes and, for this reason, are usually located near the most important cities. Maritime container terminals form a portion of a port, and the largest ones are obviously located inside the largest ports by size and trade. Inland container terminals tend to be located near large cities, with excellent rail links to the maritime container terminals. The main structures and features of a container terminal can be identified starting from the normal subdivision of its surface into 4 zones:

- The first zone is the terminal entrance (gate), where containers are checked and registered, as well as the administrative and customs procedures for inbound and outbound flows. Within this first area, there are generally administrative offices for the police, Customs, and a control tower.

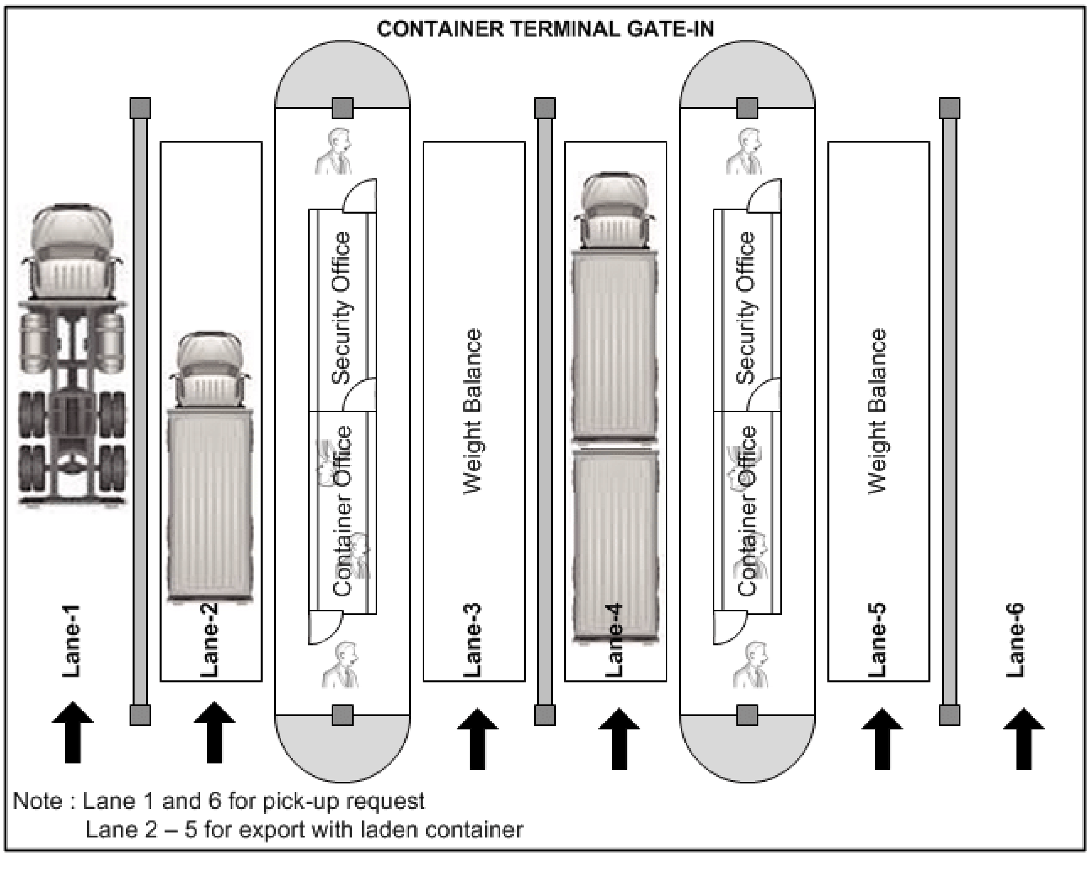


Figure 14 –Container terminal gate-in, https://www.researchgate.net/figure/Layout-of-container-terminal-gate-in_fig1_316959734

- The second zone is configured as a handling area for the rail wagons or the trucks that must receive or deliver the containers. It is characterized by parking areas for trailers and tractors. The rail lines and gates allow the circulation of rail shuttles in and out; the roads make it possible to reach the freight warehouses where the containers are positioned for loading or unloading of cargo. Nearby it is easy to find garages and maintenance workshops. The carriageways that divide the various work areas are fairly wide so as to facilitate maneuvers for the vehicles and lead to the entry and exit gates.

- The third zone is generally an area for container storage, divided into rows according to a necessary distribution scheme that identifies the position of the containers. In a rough way, two storage methods can be distinguished: in the first, the containers are placed on frames so that each one has direct access; in the second, the containers are stacked and piled on the ground and consequently are not all directly accessible. Nowadays, because of limited space, the second storage method prevails, while the frame method can be found in North America.

Most terminals organize the yard into different blocks based on the characteristics of the containers (size, reefers, empties) and on the particular cycle to which they belong, that is, import or export. A large yard is able to hold up to 20,000 TEU. The efficiency of stacking operations is determined by strategic decisions related to the choice of handling equipment used, the layout of the yard blocks, operational assessments concerning storage, and the scheduling and routing of stacking equipment. These decisions are made according to the space available, the expected container volume, the dwell time set for the containers, and regulations from the authorities.

- The fourth zone is the quay operating area. Here, containers are moved with special equipment: the so-called ship-to-shore container cranes. These cranes, aided by the standardization of dimensions and coupling systems, can deliver high operational handling speeds. Terminals use several types of cranes: rail-mounted yard cranes and rubber-tyred ones.

Before the container loading/unloading process begins, the ship must berth alongside the quay. Almost always it is assigned a berth before its arrival. Ships that belong to the same liner service are usually berthed at the same spot on the quay. After defining the berth, even before the ship arrives, the total moves must be known (number of containers to discharge/load/reposition) and their distribution on the ship, in order to define the optimal number of quay cranes needed to allow loading/unloading operations.

The rules of the International Standardization Organization define the minimum strength requirements and the dimensional characteristics of containers in order to guarantee the modularity of the system and the possibility of transport on the different modes. Their structural strength allows containers to be stacked up to four units (and beyond) in the storage yards. The elements in the container structure that allow gripping by the different lifting devices are the corner castings, into which special hooks enter to ensure the coupling between the lifting device and the container, as well as the locking of the container on vehicles. All terminal equipment is designed and assembled to move containers on and off ships and position them in the storage yards while awaiting shipment or pickup.

Space is the most useful "product" in a terminal: vast berthing areas are required where ships can lie alongside. Containers are waterproof and therefore do not need to be stored in covered areas such as sheds or quay warehouses. For customs checks, however, specific covered zones exist. Containers are stacked and divided into two different zones: in the first, export containers are deposited, which will then be loaded onto the containership; in the second, import containers are deposited. The latter will then be stowed on a feeder ship that will transfer them to another port, or they will be loaded onto trucks or rail shuttles to reach the various destinations. Container handling is performed with special vehicles which, aided by the standardization of dimensions and coupling systems, can achieve high lifting and travel speeds. Two fundamental loading techniques can be used:

- Horizontal loading, according to which the loading units are transshipped without causing them to be lifted off the ground, but using inclined platforms and/or push tractors;
- Vertical loading, according to which the loading unit (semi-trailer, swap body, and container) is transferred from the road vehicle to the rail vehicle or vice versa with cranes or special lift trucks.

Chapter 3 – Research objective and methodological approach

3.1 Research objective

The goal of this thesis is to identify inefficiencies in the logistics flow of a food supply chain by applying unsupervised clustering techniques to an operational dataset. The analysis was performed on real operational data from Ferrero that were modified and anonymized to comply with data protection policies.

This analysis focuses on the products' storage behavior, lead times, and distribution performance. To do so, products that share similar logistical patterns were analyzed using a Python algorithm

Clustering algorithms such as K-Means and DBSCAN, the study aims to uncover hidden patterns in the data, clustering products that behave similarly in warehousing time, shipping delays, expiration risk, and volume.

DBSCAN is one of the most used clustering algorithms. It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed and marks as outliers points that lie alone in low-density regions [26].

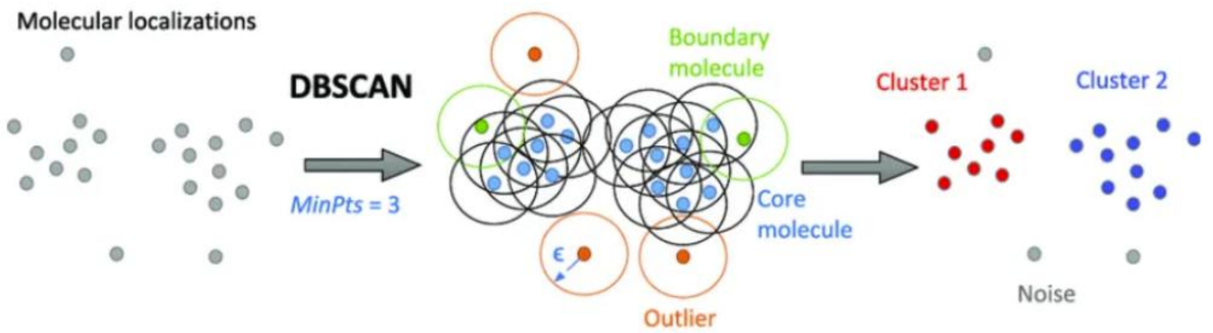


Figure 15 – DBSCAN applied to an example of molecular localizations [27]

K-Means, on the other hand, is a centroid-based algorithm that partitions data into a predefined number oif clusters (k clusters) based on the mean distance between points and their assigned centroid. The algorithm aims to minimize the sum of squared distances between each point and its assigned centroid. K-Means is widely used due to its simplicity and efficiency. [28]

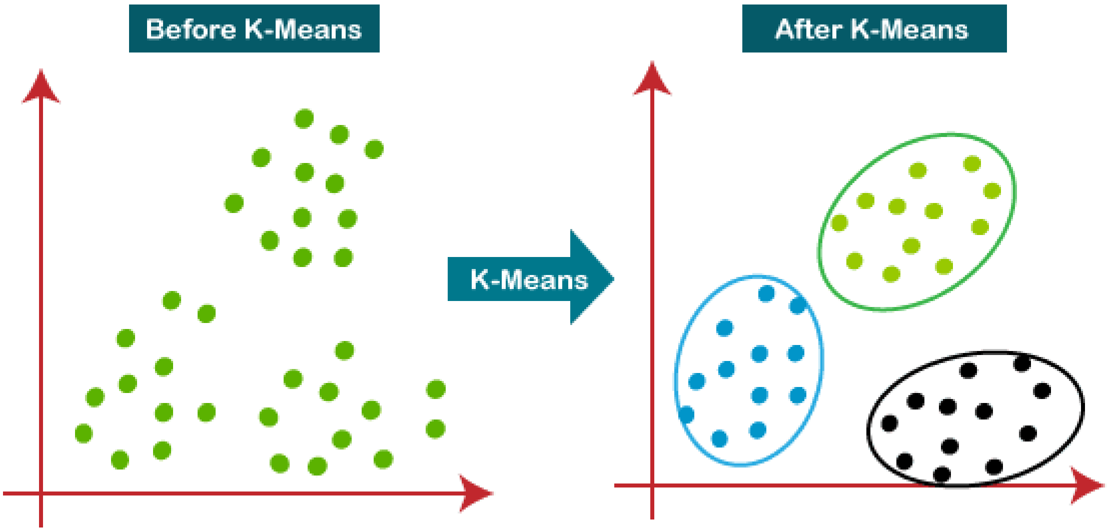


Figure 16 – K-Means applied to an example of molecular localizations [29]

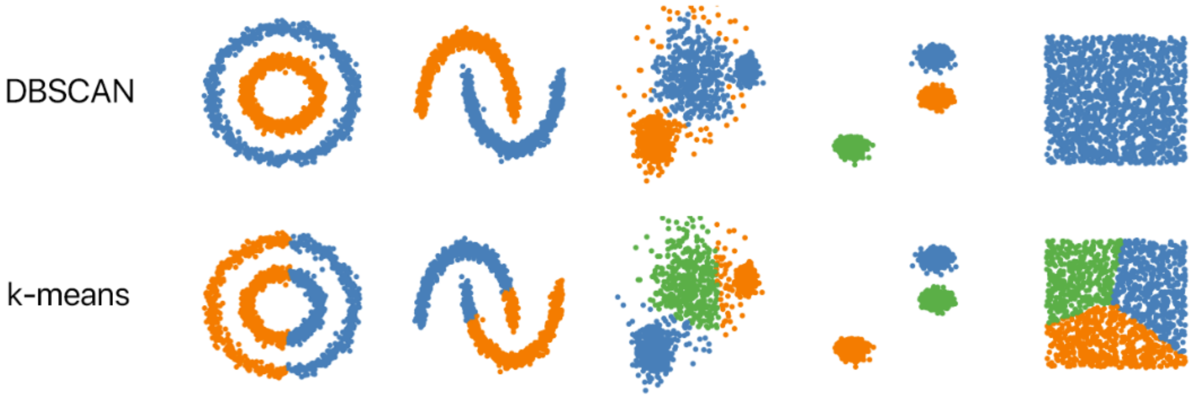


Figure 17 - Example of how different datasets are clustered by the algorithms. [30]

In the first case, with concentric cricles, K-means clustered the points based on the distance. However, the distance is not the key here. The key here is the density of the pattern. DBSCAN will be able to capture that. Thus, it is very important to do a preliminary analysis of the data before choosing the right clustering algorithm.

DBSCAN is based on the idea that a cluster in data space is contiguous region of high point density, seperated from other such clusters by contihuous regions of low point density. It finds core samples of high density and expands clusters from them.

Although both algorithms have limitations, and clustering is inherently unsupervised and not predictive, it allows for insights by revealing patterns that can anticipate future inefficiencies. For instance, identifying clusters of products that systematically experience long storage times or inefficient shipment intervals can inform strategic actions such as revising warehouse policies or optimizing stock rotation rules.

In this context, data-driven segmentation becomes a valuable tool to go deeper in the study of the SCM analysis, understanding how subsets or subclasses of products contribute to the overall system.

3.2 Methodological approach

The analysis followed a clear, sequential workflow that moved from a raw dataset to cluster-based insights.

1) the data was prepared to be analysed: the file was imported, missing values were handled, and text fields were converted into numerical format.

2) New variables, such as lead time (days between production and shipment), were derived, and key performance indicators, including the Storage Cost Index, Stock Rotation Efficiency, and Expiration Risk, were calculated. These metrics were used to assess each product's logistics performance.

3) Once the data were clean and consistent, the most relevant variables were selected and standardized to ensure comparability across scales.

4) Two unsupervised machine learning algorithms were then applied for clustering: K-Means, which partitions the data into a predefined number of groups with similar characteristics, and DBSCAN, which detects outliers and observations with unusual logistic patterns.

5) Principal Component Analysis (PCA) was used to reduce dimensionality and support visualization in order to make the data easier to visualize.

6) The resulting clusters were examined by computing average KPIs and describing the typical features of each group.

7) Visual tools such as scatter plots, heatmaps, and distribution charts were produced to aid interpretation and highlight potential logistical inefficiencies.

8) Based on these results, practical recommendations were formulated to improve warehouse rotation, container planning, and overall supply chain performance.

Chapter 4 – Ferrero case

4.1 The supply chain at Ferrero

Ferrero is a multinational company operating in the confectionery sector, founded in 1942 thanks to the idea of Pietro Ferrero. Over the years the company has grown while always remaining entirely owned by the Ferrero family, which has actively participated in the development of the business by introducing new and innovative products, first among all Nutella, introduced on the market in 1964. The company was born in Alba, starting from 1956 with the first plant in Germany, and then began its expansion run in the world market. It is currently present globally with 94 companies, 25 production plants, and through authorized resellers in 170 countries.

The Ferrero Group closed the financial year ended 31 August 2024 with consolidated turnover of €18.4 billion and is the third-largest chocolate company worldwide. [31] The company's goal has always been to study specific local strategies to closely satisfy consumers' needs. The main segment in which the company operates is that of sweets and snacks to be consumed outside meals; part of this area is the market for pralines, spreads, snacks, baked products, and icecreams.

Supply chain management has its own financial and human resources managers, who report at central level. Every production plant and distribution center has a logistics manager who coordinates the flow of material inside the facility and plans its trips; in the first case this manager reports to the plant director, but at a functional level is placed under supply chain management.

In addition, in the offices at the Alba headquarters the three characteristic functions of Ferrero's supply chain are located: industrial logistics, customer service and commercial logistics.

In industrial logistics, all the flows of finished product departing from the production plant are handled; from here, shipments directed to the customer or to the warehouse are managed. The logistics that concern the procurement of raw materials are also handled in this corporate branch. Part of customer service is all those activities that are focused on customer service and not only. More broadly, this branch can also be called service &

logistics account management. Functions such as order management, demand forecasting and flow piloting are included in this context; the latter deals with sorting the goods to the different distribution centers. Commercial logistics mainly deals with the distribution of the finished product starting from the warehouses or from the transit points and then reaching the final customer.

The Suppliers portal lists the "Delivery and Tracking" applications for the plants perimeter and provides receiving hours for many facilities across Europe, North America, Latin America, Africa, and Asia. [32]

Region	**Example sites listed on the Suppliers portal**	**Note**
Europe	Alba, Pozzuolo, Sant'Angelo, Balvano, Stadtallendorf, Villers, Arlon, Belsk	Core manufacturing footprint and EU distribution
North America	Brantford, Bloomington, Franklin Park	Receiving hours published for multiple sites
Latin America	San José Iturbide, La Pastora, Quito, Alto Camarico, Pocos	Multiple origins and receivers
EMEA/Asia	Alfreton, Cork, Manisa, Baramati, Walkerville, Hangzhou, Lightow	Broad multi-regional coverage

Table 2 - Ferrero global sites with published receiving hours on the Suppliers portal.

In recent years, Ferrero has also focused on improving sustainability across its operations. The group has made progress in reducing CO2 emissions, using renewable energy, improving packaging, and working closely with local suppliers and farmers (especially for cocoa and hazelnuts) to guarantee fair trade and environmental respect.

Ferrero wants to reduce its climate impact. By 2030 it aims to cut emissions from its own sites by 50% and cut total emissions by 43% per tonne of products, it has already cut Scope 1 and 2 by 21.7% from the 2017/18 base year. Most electricity used in factories and warehouses now comes from renewables. The company is also making packaging easier to recycle and buys key ingredients, including palm oil, to avoid deforestation. The Science Based Targets initiative has approved its climate targets. [33]

In parallel, the company strengthens transparency and shared accountability across the supply chain through a due diligence approach, structured data collection, action plans and regular progress reporting. It aims for 100 percent traceability of dairy ingredients back to the farm level, supported by audits and traceability checks. An open grievance mechanism, including the Integrity Helpline operated by a third party, ensures concerns can be raised and addressed, while public annual updates and dashboards communicate progress to stakeholders.

In conclusion, Ferrero is not only a symbol of Italian excellence but also a global company with a strong and sophisticated logistics structure. Understanding how its supply chain works is crucial for identifying opportunities for improvement, especially in terms of efficiency, timing, and storage management, which will be the focus of the following sections. It manages outbound flows to multiple regions through a huge network of manufacturing and receiving sites. To improve efficiency, some products are shipped as semi-finished products to the destination plant, where they are worked into finished products.

4.2 Outbound Planning

The outbound logistics planning process at Ferrero involves internal coordination with external parties, such as *external suppliers* and *transport/logistic providers* like in figure 18, supplemented by digital solutions tools like SAP and Blue Yonder.

The initial step here is a daily stock extraction from SAP, where the system identifies available and matured products that are ready for shipment (*stock analysis* in the image below). These products are subsequently grouped by destination and temperature needs, necessary for keeping the store in good condition, and particularly important for temperature-sensitive goods.

After grouping, a weekly transport plan is developed employing tools designed for supply chain optimization such as Blue Yonder. This plan details the products that will be shipped, where they will be sent and when they will need to be removed from the warehouse to meet the delivery timelines. Next, the request to book a carrier is sent through a suitable channel (*Plan transports* in the image). Once the booking is confirmed by the carrier, a trucking service is assigned to pick up the container from the warehouse.

The container is then transported by truck to the port, where it undergoes customs inspection and is loaded onto the ship. From there, it travels via sea transport to its international destination. Once it arrives at the destination port, the container goes through customs clearance, and finally, the goods are delivered to the local distribution center for further handling and delivery to the final customers. Throughout the entire process, different actors are involved:

- Ferrero Planners, who manage stock visibility and planning.
- SAP and Blue Yonder, which provide data and planning support.
- Carriers and Trucking Partners are responsible for transport.
- Port Operators and Customs, who handle inspections and international logistics.
- Local Distributors, who receive and manage the final delivery.

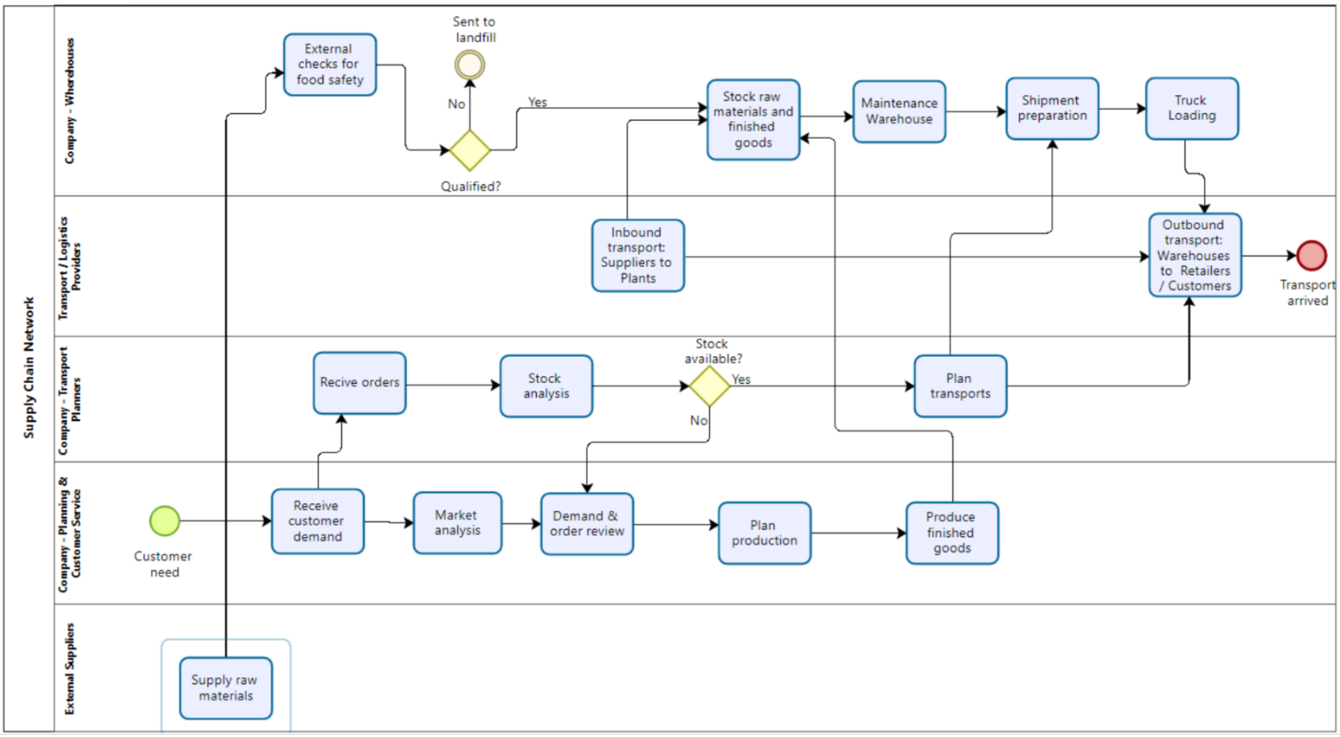


Figure 18 – Supply chain network, graph generated with Bizagi

4.3 Criticalities in Planning

Although Ferrero’s outbound logistics process is supported by clear workflows and advanced digital systems, several critical issues can still emerge along the supply chain. These issues are not always visible at first, but they have a direct impact on efficiency, product freshness, and customer satisfaction. One of the main problems is related to lead time variability. While some products are shipped just a few days after production, others remain in storage for weeks, sometimes over 30 days. This inconsistency can have implications for product quality and may lead to greater waste or product expiry. Additionally, long and volatile lead times complicate container planning, particularly when shipping abroad.

Coordination among different actors is of utmost importance. While automation and decision-making tools such as Blue Yonder support planning, this can only be achieved through communication between different departments (production, logistics, transport, etc). When this coordination breaks down or slows down, it can create delays, booking problems, or even stock imbalances. The third challenge is stock management. In some cases, products ready for shipment accumulate in warehouses, occupying space and generating extra storage costs. This occurs when the shipping slots are not synchronized with the production schedule or when containers are not available on time. This underscores the need for improved real-time monitoring and adaptable planning systems. In summary, these criticalities indicate that having a structured outbound flow is insufficient. To achieve optimal performance, the company must work on reducing variability, improving coordination, and enhancing data-driven decisions. The following chapters will explore these aspects further using clustering techniques and data analysis.

Chapter 5 – Data and processing

5.1 Dataset overview

The dataset used in this project was extracted from Ferrero’s internal systems and contains all the outbound flows connected to containerized shipments. It consists of around 70,000 rows and more than 20 columns, each representing a batch of products planned for international delivery. The dataset was modified to anonymize data and to comply with Ferrero’s data protection policies.

Some of the most relevant columns include:

- Product name and type
- Production date, maturation date, and shipping date
- Destination country
- Weight, volume, and temperature requirements
- Storage days, aging stock, and other logistics indicators

Each record relates to a product that is handled in one warehouse and assigned to one shipment. This structuring allows you to trace every single item from production to delivery from start to finish.

This dataset is targeted at studying the time lag occurring between production and shipment, and how it affects the general efficiency of Ferrero’s outbound planning. Given such a large and highly detailed dataset, it is conceivable to describe not only present operations, but to also detect hidden inefficiencies that often go unnoticed in day-to-day activities.

This dataset is a solid base for advanced analytics. Products and shipments can also be grouped according to their logistic behavior, showing the underlying patterns and the inefficiency clusters. In subsequent parts of this thesis unsupervised learning methods are used to take this raw operational data and to make it actionable while revealing space for improvement in storage, shipping, and container planning.

5.2 Description of Key Variables

The dataset includes more than 20 variables, but some of them play a key role in analyzing the efficiency of Ferrero's outbound logistics.

In this section are described the most relevant characteristics of the products that were used in the clustering phase and in the calculation of performance indicators.

In the table below are reported some of the most important columns, to give an idea of the dataset. This cluster groups lots identified by specific material codes, mainly export shipments from Italy and Poland to extra-EU markets (Mexico, Canada, United States, United Arab Emirates, Hong Kong, Senegal).

Material code	Testo breve materiale	Data Spedizione Disp.	Nazione di Partenza	Nazione di Destinazione	Calcolo Data PROD	Calcolo Data MATU
76883285	BUE T2X30X4	09/01/2024	Italia	Mexico	02/01/2024	03/01/2024
77567096	ROC TAV	10/01/2024	Polonia	Canada	02/01/2024	04/01/2024
77215103	NU G180X8X2	17/01/2024	Italia	Senegal	02/01/2024	05/01/2024
77267716	H HIP CA T1X28X6	07/02/2024	Italia	Arab Emir.	02/01/2024	04/01/2024
77242546	K JOY T16X20	31/01/2024	Polonia	United States	03/01/2024	24/01/2024
77248833	RAF T1X96X6	04/09/2023	Polonia	Hong Kong	10/08/2023	31/08/2023

Table 3 – Dataset Key Columns for Export Shipments

In This section are reported the main variables in the dataset:

Destination

The destination variable indicates the country to which the product batch is being shipped. Since Ferrero is an international company, this column includes a wide range of countries from different continents. Understanding the destination is important because it can affect lead time, shipping method, and temperature requirements. For example, products shipped to countries with long distances or complex customs procedures may have higher risks of delays or inefficiencies.

Product Type

The product type, also known as the material code, identifies the specific kind of product in each row. This includes items like chocolates, snacks, or spreads, each with different logistic characteristics. Some products require cold temperatures, others have short shelf lives, and others may be produced or packaged in different ways. This variable was encoded during the preprocessing phase to make it usable for clustering analysis.

Production and Shipment Dates

Production and Shipment Dates are both important numbers we can use to determine lead time (days from time of production till time of shipment). A long lead time can result in a product being stored in a warehouse for too long, which is expensive in terms of storage and may also increase the risk of expiration. These two dates were cleaned and renamed in the dataset to ensure consistency in all operations.

Weight and Volume

This variable reports the weight of each shipment lot, measured in quintals. Weight is employed for KPI computation – storage cost. Heavier lots generally require more work to handle and higher storage exposure. In cases where the weight value failed to appear in the source file, a preprocessing procedure dealt with these instances to avoid biased analysis. Since the project is dealing with containerized flows, weight needs to be handled jointly with saturation, that is how much of the physical limit of a container is actually used. Two constraints matter:

- Payload weight constraint. Every container type has a maximum payload (cargo weight excluding the tare). For reference, Maersk lists typical max payloads of roughly 28.3 t (for a 20′ dry), 28.9 t (for a 40′ dry) and 28.7 t (for a 40′ high-cube); the internal cubic capacities are roughly 33 m^3, 67–68 m^3, and 76 m^3, respectively. Reefer containers, with insulation and machinery, have lower payloads. These limits can also vary by country and road rules, which means operational planning would need to consider the particular lane. Reference values for weight and volume limits have been taken from Maersk's official documents, an ocean carrier commonly used by Ferrero in the shipments examined.

- Cube (volume) constraint. Even when payload is below the limit, a shipment can still "fill the box" by volume if products are bulky or low-density. In practice, the binding constraint is the smaller of weight utilization and volume utilization.

In this thesis, "saturation" is interpreted as:

- Weight utilization = total cargo weight / max payload of the assigned container type.

- Volume utilization = estimated cargo volume / internal cubic capacity.

As a matter of operational principle, compliance also requires the Verified Gross Mass (VGM) (the total of cargo plus packaging/bracing plus container tare) to be supplied before loading, pursuant to the provisions of SOLAS and acceptable in the market for carriers. It guarantees safe stowing and compliance with standard weight regulations for both sea and road legs. Connecting quintals to saturation provides the analysis with the ability to discriminate weight-constrained vs. cube-constrained types of patterns and the recognition of the pattern of product density, packaging, or container choice which leads to a systematic degradation of efficiency.

Product Type

In this dataset each product batch is assigned to a temperature class that governs storage and transport conditions and, indirectly, the maturation rules and shipment timing.

- Ambient. Confections are made in cool and dry conditions (usually 15–18 °C, humidity control). In practical application, hazelnut-based preparations such as Nutella tend to be kept at about 18 °C, which stabilizes the fat phase, supports texture, and helps aroma expression without oil separation. Product goes under ambient conditions in dry equipment (non-refrigerated), with insulation or seasonal protection as appropriate.

- Hiber. "Hiber" denotes a frozen regime used to hold product in a dormant state; in this project it refers to −23 °C. At this temperature biochemical and physical changes are strongly slowed down, extending stability during long transits or buffers. Hiber flows require reefer equipment, power availability at terminals and ships, and tighter handling procedures (pre-trip inspection, temperature logs, defrost strategies at destination).

- Frigo. Frigo is also described as chilled distribution, normally around 0–5 °C, so it is used in goods that have to be refrigerated throughout the chain. It requires reefer containers and employs full cold-chain controls; payload limits are lower than for dry units, and lead times are usually more time-sensitive.

Why this matters for the analysis:

- The kinetics of maturation are different for different setpoints, so time from production to maturation and from maturation to shipment differs by class.
- Equipment choice, payload/cube constraints, and running costs vary across classes, which influence routing and container planning.
- Temperature class is included as a categorical feature in the clustering, helping to separate ambient flows with longer storage profiles from frozen (hiber) lanes designed for long ocean legs, and from chilled frigo lanes with stricter time windows.

.**Maturation Time**

In this project, maturation time is defined as the difference in days between the production date and the maturation date:

(Maturation_Days = Calcolo Data MATU – Calcolo Data PROD).

This feature is relevant because it sets the earliest possible release time for outbound planning, and it's dependent on temperature class.

5.3 Data cleaning

Before proceeding with the analysis, the dataset underwent a structured data cleaning phase aimed at improving its overall quality and usability. The initial stage involved searching for and handling missing values. Numerous factors, particularly those with a quantitative logistics indicator, had incomplete data.

In particular, the *Weight* and *Volume* fields contained a significant number of missing entries, with over 55,000 and 76,000 null values respectively. These omissions were primarily due to gaps in system registration for certain product categories or incomplete integration between operational platforms.

Given the analytical relevance of these variables used in the calculation of indicators such as Storage Cost Index and Volume Efficiency the dataset was preserved by applying a median imputation strategy. This method was selected over alternatives (such as mean imputation or deletion of rows) to ensure a robust estimation that is less sensitive to outliers and irregular distributions.

In the context of large-scale datasets such as this one, it is essential to have a programmatic approach capable of identifying and correcting data inconsistencies. Manual inspection is not scalable when dealing with tens of thousands of entries. Therefore, the preprocessing script was developed to automatically detect missing values, replace them where appropriate, and standardize the variables required for the model. This systematic checking mechanism ensures that the quality of the input data remains consistently high throughout the analysis pipeline.

Moreover, having a resilient data cleaning routine contributes significantly to the reproducibility and scalability of the analysis. In industrial settings like Ferrero’s, it’s important to have a program that is easily adaptable to any dataset, that can be updated or extended regularly. A repeatable data preparation protocol allows for future re-analyses without the need to manually repeat all steps, ensuring methodological consistency and saving time in future applications.

Sample Code from the Preprocessing Script

To make the cleaning process more transparent, a portion of the actual Python script is shown below. This part is responsible for selecting the relevant features and applying median imputation to fill in missing values before clustering:

```
# Select numeric features for clustering
    # Features numeriche per clustering
          numeric_features = [
        'Weight', 'Volume', 'Volume_Efficiency',
        'Storage_Days', 'Lead_Time', 'Aging_Stock',
        'Storage_Cost_Index', 'Stock_Rotation_Efficiency', 'Expiration_Risk'  ]


# Apply median imputation to missing values
```

```
df[features_to_impute] =
df[features_to_impute].fillna(df[features_to_impute].median())
```

This simple but effective routine ensures that the entire matrix used for unsupervised learning is complete and free from structural gaps. The choice of the median as a replacement value is based on its resistance to outliers and better alignment with real-world industrial data, which often includes extreme or skewed distributions.

In addition to managing missing values, the data cleaning phase also involved several essential preprocessing steps:

- **Encoding of categorical variables** such as product type, destination, warehouse, and temperature conditions, through *label encoding*. This technique transforms each category into a unique numerical value, allowing the data to be processed by machine learning algorithms. For example, in the case of the "Destination_Country" variable, each country is assigned a specific integer code: "Germany" → 0, "Italy" → 1, "France" → 2, "Canada" → 3, "USA" → 4.
- This encoding method preserves the categorical nature of the data without implying any ordinal relationship.

- **Normalization of numerical features** using standard *z-score scaling*, ensuring that variables with larger ranges do not dominate the clustering process. This step is particularly important when combining multiple indicators with different units and magnitudes.

- **Column renaming and semantic mapping** to enhance clarity and align the variable names with the objectives of the analysis. For example, the column originally named “GG da PRODUZIONE” was renamed as “Days_Production” to improve readability and consistency.

These operations were fundamental in transforming a raw operational dataset into a structured and machine-readable analytical base. The resulting dataset was thus ready to support the clustering and pattern recognition activities that follow in the subsequent chapters.

5.4 Lead time calculation

One of the key variables used in the efficiency analysis of Ferrero's outbound logistics is the lead time, defined as the number of days between the production date and the shipment date of each product batch. This variable, referred to as Lead_Time in the dataset, serves as a fundamental indicator of planning and warehouse performance.

A high lead time can signal potential inefficiencies such as excessive storage, poor alignment between production and shipment scheduling, or logistical bottlenecks. Conversely, a very short lead time might indicate a just-in-time strategy but could also reflect an overstrained distribution chain with limited buffer capacity.

For this reason, it was necessary to calculate this feature explicitly, as it did not appear in the original dataset. The calculation was carried out programmatically using the Pandas library in Python, which computes the difference in days between the standardized Data_Produzione and Data_Spedizione columns. Here is the relevant code snippet used in the preprocessing phase:

```
# Calcolo del Lead Time (differenza in giorni tra produzione e spedizione)

Df ['Lead_Time'] = (df['Data_Spedizione'] - df['Data_Produzione']).dt.days
```

This simple yet crucial operation transformed two separate date fields into a meaningful metric that could be used for downstream analysis, including:

- Classification of storage efficiency.
- Clustering similar logistic behaviors.
- Identification of critical products with excessive waiting times.

The computed Lead_Time variable was also used in the creation of derived KPIs, such as Stock Rotation Efficiency, and played a key role in detecting outlier patterns and inefficient clusters.

Finally, the robustness of this transformation was ensured by checking for null values and invalid date formats before applying the operation, to avoid any inconsistencies in the results. Through this feature engineering step, the dataset became better aligned with the strategic questions posed in the analysis, enabling a data-driven approach to performance evaluation.

5.5 Performance indicator calculation

This section describes the indicators that were built from the raw variables to support the analysis and the clustering step. The goal is to convert operational data into simple measures that capture storage exposure, rotation speed, risk of obsolescence, and space use in the container flow. All indicators were computed after cleaning dates and handling missing values, as described in the previous sections.

The following key performance indicators were used:

- **Storage Cost Index**: an indicator for exposure in a warehouse. It is computed as *Storage_Days × Weight*.
- **Stock Rotation Efficiency**: is the measure of how effectively a company cycles through its inventory, ensuring older stock is sold and replaced quickly, thereby reducing waste and costs. It is computed as *1 / (Lead_Time + 1)*. Adding one avoids division by zero for same day releases.
- **Expiration Risk**: a simple ratio that highlights products with high aging against time in storage. It is computed as *Aging_Stock / (Storage_Days + 1)*.
- **Volume Efficiency**: an indicator of space use. It is computed as *Weight / Volume*. If volume is in cubic meters and weight in quintals, conversions can be applied to obtain a consistent scale. This ratio helps to distinguish weight-constrained from cube-constrained patterns.

To keep the indicators robust, divisions handle zeros with a small offset, and extreme values are clipped to reasonable ranges. The code below shows the exact implementation used for the dataset.

Here is the function of the program that converts raw variables into key performance indicators.

```
def calculate_supply_chain_kpis(df):

  # Storage Cost Index
  if 'Storage_Days' in df.columns and 'Weight' in df.columns:
    df['Storage_Cost_Index'] = df['Storage_Days'] * df['Weight']
    print("Calcolato Storage_Cost_Index = Storage_Days * Weight")

  # Stock Rotation Efficiency
  if 'Lead_Time' in df.columns:
    df['Stock_Rotation_Efficiency'] = 1 / (df['Lead_Time'] + 1)
    print("Calcolato Stock_Rotation_Efficiency = 1/(Lead_Time + 1)")

  # Expiration Risk
  if 'Aging_Stock' in df.columns and 'Storage_Days' in df.columns:
    df['Expiration_Risk'] = df['Aging_Stock'] / (df['Storage_Days'] + 1)
    df['Expiration_Risk'] = df['Expiration_Risk'].clip(0, 10)  # Cap outliers
    print("Calcolato Expiration_Risk = Aging_Stock / (Storage_Days + 1)")

  # Volume Efficiency
  if 'Weight' in df.columns and 'Volume' in df.columns:
    df['Volume_Efficiency'] = df['Weight'] / (df['Volume'] + 0.001)  # Evita div by zero
    print("Calcolato Volume_Efficiency = Weight / Volume")
```

5.6 Inefficiency Score

The composite median imputationsummarizes three drivers of outbound inefficiency for each shipment record: days in storage, production-to-shipment lead time, and aging stock.

Each driver is first converted into a binary flag that indicates whether the value is unusually high relative to the dataset. The three flags are then summed to obtain a score from 0 to 3, which is finally mapped to an efficiency class.

How the "high" threshold is defined.
For each variable X(*Storage_Days, Lead_Time, Aging_Stock*) the threshold is set at the 75th percentile, also called the third quartile Q_3. By definition, 75 percent of observations are at or below $Q_3$and the top 25 percent are above it. Using Q_3makes the rule adaptive to the data distribution and less sensitive to outliers than a mean-based rule.

Formally, the three flags are:

- **High_Storage** = *1 if Storage_Days >* Q_3*(Storage_Days), else 0*
- **Long_Lead_Time** = *1 if Lead_Time >* Q_3*(Lead_Time), else 0*
- **High_Aging** = *1 if Aging_Stock >* Q_3*(Aging_Stock), else 0*

The composite score is:

$$\boldsymbol{Inefficiency\_Score} = High\_Storage + Long\_Lead\_Time + High\_Aging \in \{0,1,2,3\}.$$

The efficiency class is assigned as:

- 0 → Efficient
- 1 → Moderate
- 2 → Inefficient
- 3 → Critical

This approach produces an interpretable, scale-free indicator. It highlights the quartile of records that simultaneously show long waits in storage, long release times, and high aging exposure, without fixing arbitrary business cut-offs.

5.7 Feature Preparation for Clustering

This step transforms the cleaned dataset into a model-ready matrix for K-Means and DBSCAN. In the script, this is implemented in the function that prepares features for clustering. The logic is simple: select the most informative numerical variables, encode the key categorical fields, and combine everything into a single feature set.

Numerical variables:

The program collects the main logistics metrics that describe storage exposure, timing, and space use: *Storage_Days, Lead_Time, Aging_Stock, Weight, Volume, Volume_Efficiency, Storage_Cost_Index, Stock_Rotation_Efficiency, and Expiration_Risk*. These were chosen because they affect consolidation, warehouse rotation, and container planning. Units are kept consistent with the earlier cleaning steps, and any residual gaps are handled later during imputation in the clustering phase.

Categorical variables:

The program then converts the categorical fields into numeric codes using label encoding. In practice, each distinct category is mapped to an integer identifier without introducing any order. For example, Destination_Country might map Italy→0, Canada→1, United States→2, and so on. The encoded fields include Product_Type, Division, Warehouse, Destination_Country, Strategy, and Temperature. In the current dataset the cardinalities are high for Product_Type (about 1,400 categories) and moderate for the others (Division ≈ 7, Warehouse ≈ 36, Destination_Country ≈ 69, Strategy ≈ 3, Temperature ≈ 3). Label encoding keeps the feature space compact, which is helpful for runtime and memory with K-Means and DBSCAN. This choice is noted in the discussion, since Euclidean distance treats these codes as numeric; the effect is mitigated by standardization in the next step and by using only a small set of categorical fields that act as profile indicators.

Feature assembly. After selecting the numerical variables and creating the encoded versions of the categorical ones, the program concatenates them into a single DataFrame and stores the list of feature names. This matrix is the direct input to scaling and clustering.

The code below implements the procedure described above:

```
# Numerical features

num_features = [

    'Storage_Days', 'Lead_Time', 'Aging_Stock',
    'Weight', 'Volume',
    'Volume_Efficiency', 'Storage_Cost_Index',
    'Stock_Rotation_Efficiency', 'Expiration_Risk']
available_num = [c for c in num_features if c in df.columns]
X = df[available_num].copy()

# Categorical features to encode
cat_features = ['Product_Type', 'Division', 'Warehouse',
                'Destination_Country', 'Strategy', 'Temperature']
for col in cat_features:
    if col in df.columns:
        le = LabelEncoder()
        df[col] = df[col].fillna('Unknown')
        df[col + '_encoded'] = le.fit_transform(df[col])
        X[col + '_encoded'] = df[col + '_encoded']

# At this point X is the feature matrix used for scaling and clustering
feature_list = X.columns.tolist()
```

This procedure produces a compact and interpretable set of features that capture timing, exposure, density, geography, and temperature class. In the following step, the matrix is standardized and used to fit K-Means and DBSCAN, with median imputation applied where needed.

Chapter 6 – Clustering techniques

6.1 Machine learning techniques in food logistics

In recent years, machine learning has become an increasingly important tool in logistics and supply chain management. Thanks to its ability to work with large amounts of data, it helps companies understand patterns and make better decisions in complex situations, especially when things change quickly, like food or container transport.

The food industry, represents a particularly challenging case as the supply chain management for short life cycle products is a significant issue for companies dealing with perishable goods. For definition, a good that has a short life cycle is produced and sold only for a limited period of time typically under 12 months [34]

Due to the perishable nature of the goods, short-life cycle products require a more responsive and agile organisation compared to long-life ones.

As stated by M.L. Fisher [35], products with short life cycles require a supply chain that is completely different from the one suitable for standard products. Supply chain managers must be able to face the effect of an unstable demand and to learn to adjust it, taking into account to the continuous changes in customers' needs and requirements

Some of the most common uses of machine learning in logistics are:

- Forecasting product demand
- Optimizing warehouse inventory
- Planning routes and deliveries
- Predicting maintenance needs
- Detecting delays or problems in operations
- Grouping similar products or shipments

This project is focused on unsupervised learning, i.e., the dataset is unlabelled where the goal is to group products with similar logistic behavior, based on characteristics like lead time, weight, storage days, and expiration risk. Clustering analysis is performed by computing size, mean KPIs, efficiency distribution, and enumerating main products and destination countries so that every product group has a defined profile. Inefficiencies are highlighted by comparing the clusters to see which are critical vs efficient and by recognizing common patterns: many storage days, long lead times, expiry risk, etc.

A dashboard is developed to display the results quickly, like the cluster distribution, the average inefficiency, a chart showing storage days against lead time, and a KPI heatmap to make it instantly easy to understand. Then turningg these results in a strategic recommendations: for critical clusters, like reducing storage days, improve lead times. Another important solution is the review of the warehouse strategy to plan better when seasonal peaks arise. Additional analyses look at how much of a relationship regular shipment frequency has with stock accumulation, and study cost patterns across clusters.

The two main algorithms used are:

- **K-Means**: This algorithm divides the data into a number of groups (clusters) that have similar characteristics. It is simple and works well when we can estimate the number of groups. It helps to find different types of product profiles, based on their logistic performance.

- **DBSCAN**: This is another clustering algorithm, but it doesn’t need to know the number of groups in advance. It is very useful to find outliers, which are products with unusual behavior that might be inefficient or risky.

To increase the conmputatoional efficiency, data preparation techniques were employed. For example,numbers were normalized and texts were converted into categories. Then, PCA (Principal Component Analysis) was used to reduce the number of variables and make the results easier to visualize.

Thanks to these tools, it's possible to go from just observing problems form the dataset to predicting them. In the case of the Ferrero supply chain, this approach helps to analyze how products are stored and shipped, how containers move, and where improvements can be made.

6.2 Introduction to clustering

Clustering is an unsupervised learning technique that groups records which are similar to each other and dissimilar from the rest, without using labels or targets. In logistics, clustering is useful when large operational datasets hide recurring patterns that are not obvious a priori: it can reveal typical shipment behaviours, isolate atypical or risky cases, and provide segments on which to design targeted actions.

In this study, clustering serves three purposes: to detect macro-patterns in storage and release timing, to separate a small set of shipments with unusually high exposure to delays or aging, and to create a quantitative base for decisions on warehouse rotation and container planning. Two complementary methods are adopted.

K-Means produces a global partition of the dataset into a small number of compact groups, offering a clear overview of the main behaviours. DBSCAN focuses on density, discovering small dense structures and flagging sparse points as potential anomalies; this is valuable for exception management.

Before modelling, the dataset was cleaned and mapped (about 78,000 rows, more than 20 columns). Core fields include production, maturation and shipment dates; storage days; aging stock; weight and volume; product, warehouse, destination and temperature. Lead time was computed as the difference between shipment and production dates.

From these, several KPIs were derived to characterise logistics performance:

- Storage Cost Index = Storage_Days × Weight;
- Stock Rotation Efficiency = 1/(Lead_Time + 1);
- Expiration Risk = Aging_Stock/(Storage_Days + 1);
- Volume Efficiency = Weight/Volume.

A composite inefficiency score was then built using quartile-based rules on storage, lead time and aging, and records were classified as we can see in *Image 3*, reported below with the following values: Efficient 57.5%, Moderate 19.0%, Inefficient 16.1%, Critical 7.5%.

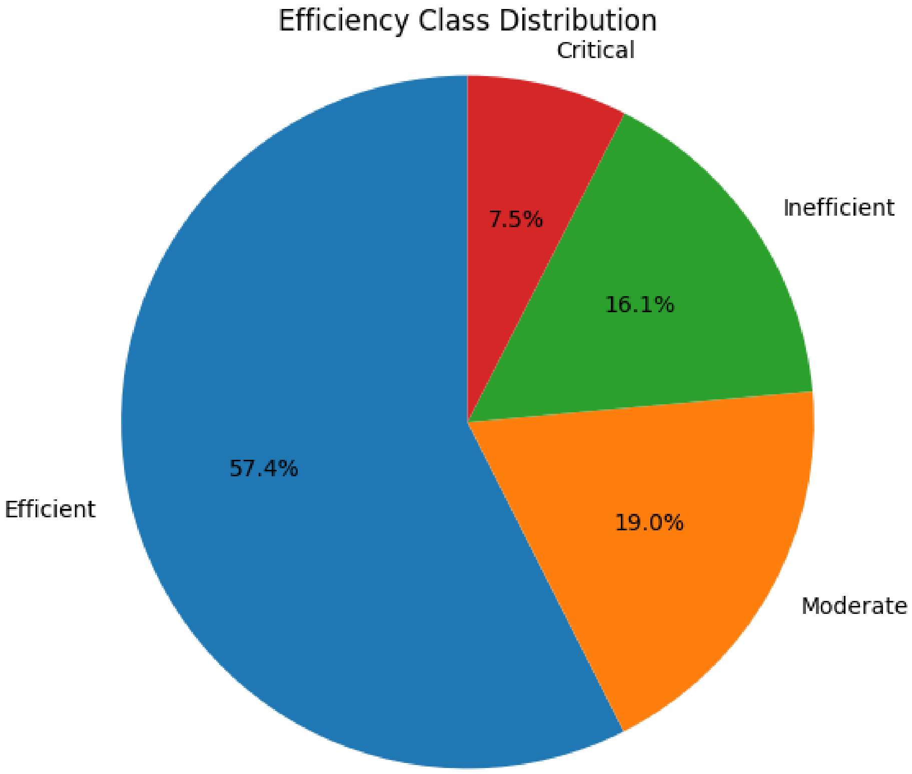


Figure 19 – Efficiency class distribution obtained using quartile, Python program output

For clustering, fifteen features were used: nine numerical metrics and six categorical variables encoded as integers. The feature matrix contained missing cells mainly in weight and volume; these were imputed with the median, and all features were standardized. This preparation enables distance-based algorithms to operate on a consistent scale and to focus on relative behavioral differences rather than unit magnitudes.

6.3 K-Means

K-Means is an unsupervised clustering method that partitions a dataset into K groups by minimizing the within-cluster sum of squares: each record is assigned to the nearest centroid (Euclidean distance), centroids are recomputed, and the process iterates until assignments stabilize.

Because Euclidean distances depend on feature scale, inputs must be standardized. In this study, the datas were first imputed with missing numeric values (using the median for weight and volume), label-encoded for key categorical fields (such as product, destination, warehouse, temperature, etc.), and then standardized using z-scores with StandardScaler. The resulting standardized matrix (X_scaled) is what K-Means and DBSCAN use, ensuring all variables contribute equally and that no single unit dominates the distance.

This unsupervised clustering method works best when clusters are compact and roughly spherical in the standardized space. It can be sensitive to outliers and to the choice of K, so we ran multiple initializations and evaluated several K values using diagnostic criteria (elbow/silhouette). K-Means minimizes the inertia, i.e., the sum of squared distances from each point to its cluster centroid. If you plot inertia vs. the number of clusters K, the curve always goes down as K grows.

The elbow method looks for the point where the curve bends: before the elbow, adding clusters gives large improvements; after the elbow, improvements are marginal. That bend suggests a good trade-off between simplicity and fit. Limitations: sometimes the curve has no sharp bend (especially with imbalanced or noisy data), so the elbow is subjective. That's why we pair it with silhouette.

The silhouette quantifies how well each point fits its assigned cluster vs. the nearest alternative cluster. For a point i:

- $a(i)$= average distance to points in its own cluster (compactness),
- $b(i)$= lowest average distance to points in the closest other cluster (separation),
- silhouette $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, which lies in $[-1,1]$.

Interpretation:

- near 1.0 → very well clustered (tight and well separated),
- around 0.5 → reasonable structure,
- around 0.0 → overlapping clusters,
- negative → likely misclassified.

Application to Ferrero data

K-Means was fitted on fifteen features: nine numerical indicators (*Storage_Days, Lead_Time, Aging_Stock, Weight, Volume)* and derived KPIs such as *[Storage_Cost_Index, Stock_Rotation_Efficiency, Expiration_Risk, Volume_Efficiency]* plus six categorical descriptors encoded as integers (*Product_Type, Division, Warehouse, Destination_Country, Strategy, Temperature*).

After imputation and standardization, we tested *K* = *2…9*. The best separation was at K = 2 with silhouette = 0.504; larger *K* values yielded lower silhouettes (≈0.14–0.19).

The best separation was obtained at K = 2 with a silhouette score of 0.504. Higher K values produced lower silhouettes (0.14–0.19), indicating weaker separation. The final split is Cluster 0 (C0) = 74,404 records, 95.1% and Cluster 1 (C1) = 3,860 records, 4.9%, see Figure 4 below.

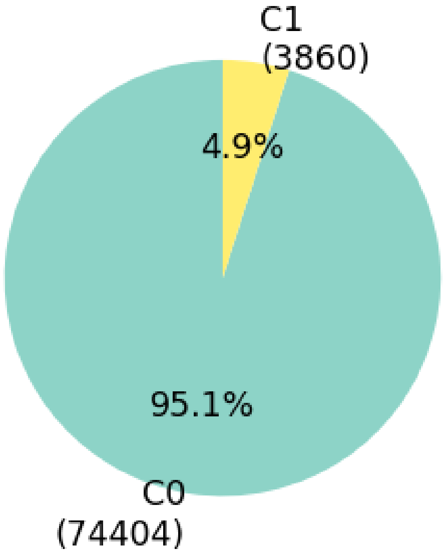


Figure 20 – Product distribution per Cluster, Python program output

With $K = 2$, K-Means separates the dataset into a dominant baseline group and a smaller, atypical group. Cluster 0 contains 74,404 records (95.1 percent) and shows typical operating values. Cluster 1 contains 3,860 records (4.9 percent) and concentrates the highest exposure on the main logistics metrics.

In particular, average storage days are about 105 in *Cluster 0* versus about 660 in *Cluster 1*; average lead time (production to shipment) is about 18 days in *Cluster 0* versus about 32 days in *Cluster 1*; average aging stock is about 148 in *Cluster 0* versus about 1,123 in *Cluster 1*. The composite inefficiency score confirms this gap, with a mean of 0.64 in *Cluster 0* and 2.49 in *Cluster 1* on a 0–3 scale.

The two clusters are also separable in a two-dimensional PCA projection, which is used only for visualization (explained variance: PC1 ≈ 20.8 percent, PC2 ≈ 14.2 percent, PC3 ≈ 9.8 percent). Scatter plots of storage days versus lead time show a positive association, with *Cluster 1* concentrated at higher values. Weight and volume plots indicate that *Cluster 1* includes heavier and bulkier items on average, which may interact with consolidation choices and capacity constraints. Together, these views support a consistent reading: a small minority of records displays prolonged storage, longer release cycles, and higher aging exposure.

Figure 21 (below) summarizes the difference in the composite inefficiency score between the two clusters and can be included as a compact diagnostic in the chapter.

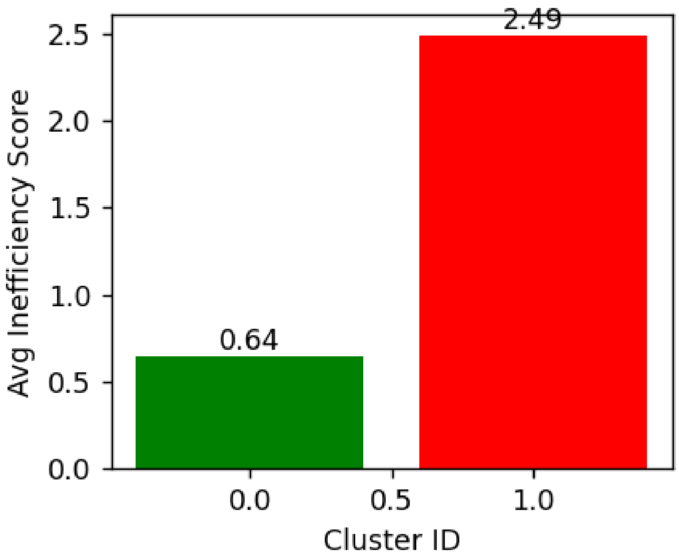


Figure 21 – K-Means results: average inefficiency score by cluster, Python program output

In Figure 22, we can analyze the Average Storage Days by Cluster.
The bar chart compares mean storage days for the two K-Means groups. Cluster 0 averages ~105 days, while Cluster 1 reaches ~660 days—about 6.3× higher. This large gap confirms that the minority cluster accumulates prolonged warehouse time, consistent with its higher aging stock and inefficiency score. Operationally, Cluster 1 should be prioritized for rotation policies and booking/scheduling checks.

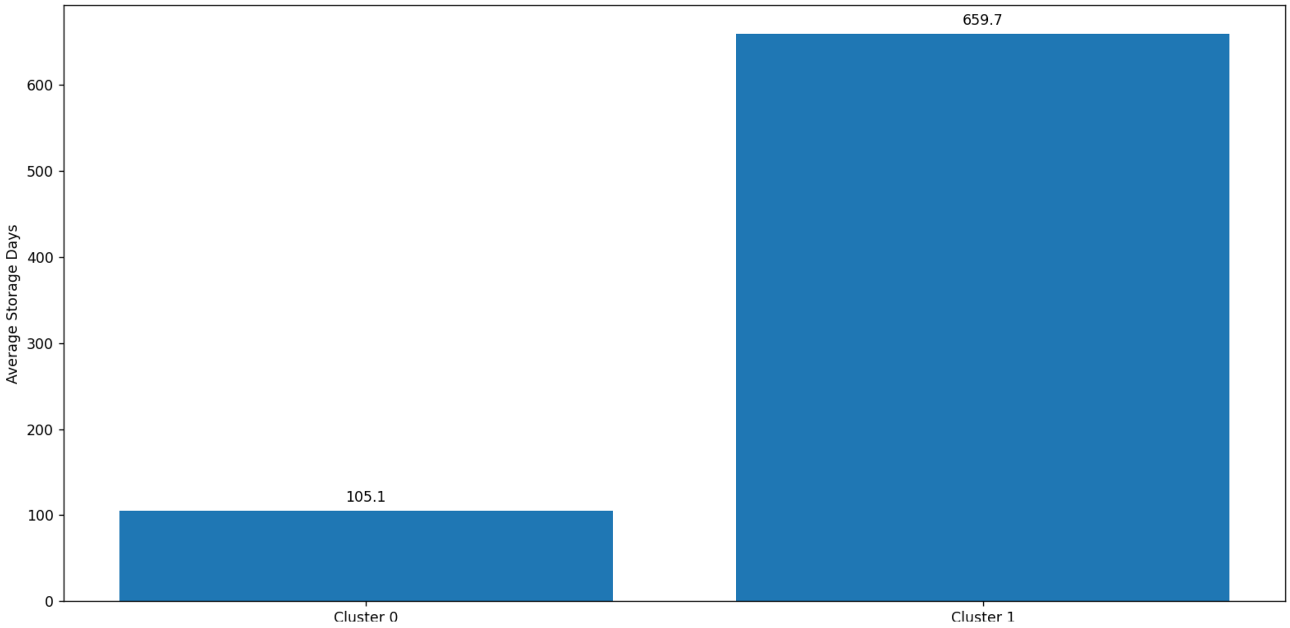


Figure 22 - Average Storage Days by Cluster, Python program output

6.4 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised method that builds clusters from regions of high local density and labels isolated records as noise.

This method groups records based on local point density rather than forcing a fixed number of clusters. Two parameters control the behaviour:

- *eps* defines the neighbourhood radius used to measure density;

- *min_samples* is the minimum number of neighbours required to form a dense region.

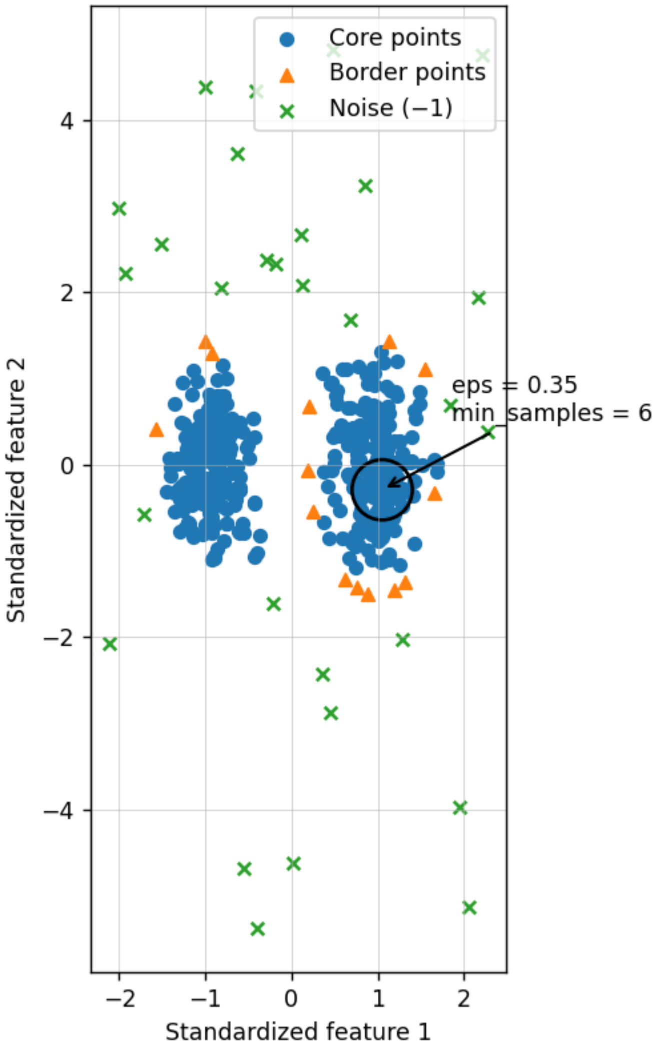


Figure 23 - Dbscan density (eps & min_samples), Python program output

To understand better how to read the figure 23

- Blue dots = core points: each has at least 6 neighbours inside the eps = 0.35 radius (the black circle shows one example), so they lie in a high-density area.

- Orange triangles = border points: not dense enough to be core, but density-reachable from a core point, so they join its cluster.

- Green X = noise/outliers (*label –1*): isolated points outside any dense region.

- The two compact clouds illustrate two clusters found by DBSCAN; axes are standardized features (z-scores) so distances are comparable.

Parameter Selection is empirical. A k-distance plot on the standardized data serves to understand a reasonable eps (which corresponds to the knee at which distances begin to increase steeply). The min_samples value is established on the basis of dimensionality, in the range of approximately 2 to 4 times the number of features, and then examined for coherence in neighboring values. High dimensional spaces, where distances may focus and differences in eps may affect cluster identity, should be carefully taken into account.

DBSCAN complements K-Means by adding a local density perspective. K-Means gives you a large macro-partition that's easy to segment by means of KPIs, while DBSCAN reveals micro-clusters and valid outliers that dwell in low-density regions. In the 15-feature matrix, it reveals small and consistent pockets, usually related to a particular destination, temperature, warehouse pair, and identifies a noisy list of shipments for focused detection. In practice, the integrated reading is handy: K-Means facilitates policy design and global monitoring, DBSCAN provides an operational exception list for root cause analysis (e.g. long storage in a lane repeatedly, or serial cold chain batches that stack up). Using a Standardized Feature Matrix, DBSCAN (with eps = 0.5 min_samples at library default) returned 713 dense clusters and 11,287 points as noise, or 14.4 percent of the sample. Put differently, the algorithm found many small, locally coherent pockets of behaviour, and a considerable dataset of atypical records. The output goes along with the K-Means segmentation. K-Means yielded a global, compact split into two macro-groups (C0 ≈ 95.1 percent C1 ≈ 4.9 percent), exhibiting a minority cluster with far higher exposure on storage, lead time and aging. DBSCAN, in contrast, identified low-level density islands and isolated cases that behave different between one point and another. In practice, both readings have varying applicability: K-Means allows for high-level profiling and policy design, while DBSCAN is used for exception tracking, and for trace back some patterns (e.g., small sets of shipments linked to a specific destination, temperature class or warehouse having repeated anomalies).

6.5 Cluster Profiling

This section profiles the K-Means partition, since it offers a compact and interpretable view for managerial use. PCA was used only for visualization; the first three components account for about 20.8 percent, 14.2 percent, and 9.8 percent of variance respectively, which is adequate to support two-dimensional plots without being used in model fitting.

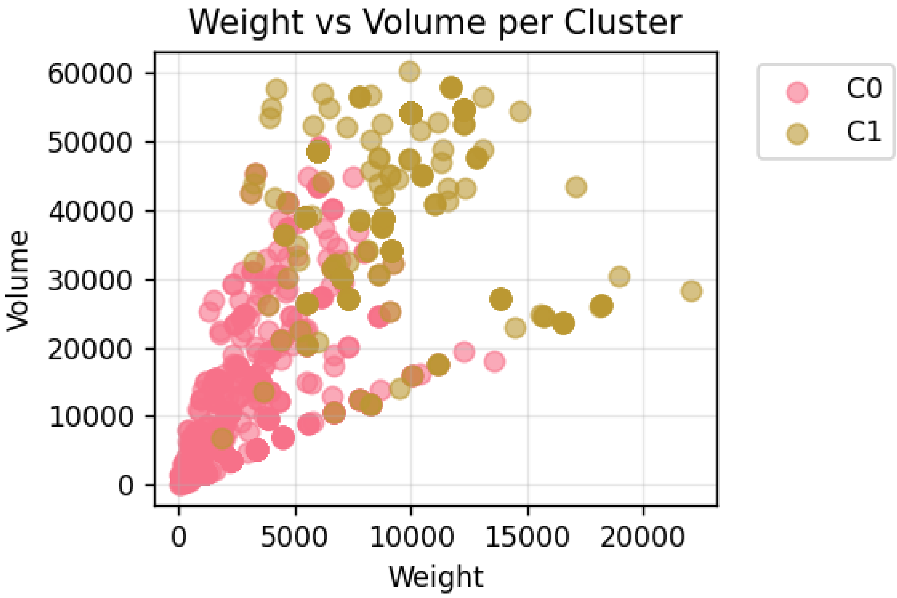


Figure 24 – Weight vs Volume per Cluster, Python program output

The chart “Weight vs Volume per Cluster” shows a clear positive relationship: as shipment weight increases, volume tends to increase as well. Cluster C0 (pink) forms a dense cloud in the lower-left quadrant, indicating many small shipments with relatively low weight and volume and a fairly compact spread. Cluster C1 (gold) extends into higher ranges on both axes and is more dispersed, which points to larger, more variable shipments.

For a similar weight, C1 often lies higher on the volume axis than C0. This suggests lower packing density for C1 shipments that are bulkier per kilogram, possibly due to packaging constraints, temperature-controlled configurations, or stacking limits. A few isolated points with very high weight or very high volume relative to the other dimension are potential outliers worth a specific review.

From an operational perspective, the plot helps identify where container utilization could be improved. Combining this view with a density metric (weight/volume) and the storage or lead-time KPIs can highlight bulky and slow-moving products or lanes that should be prioritized in packing rules, load-building strategies, or booking policies.

Distribution by destination and product

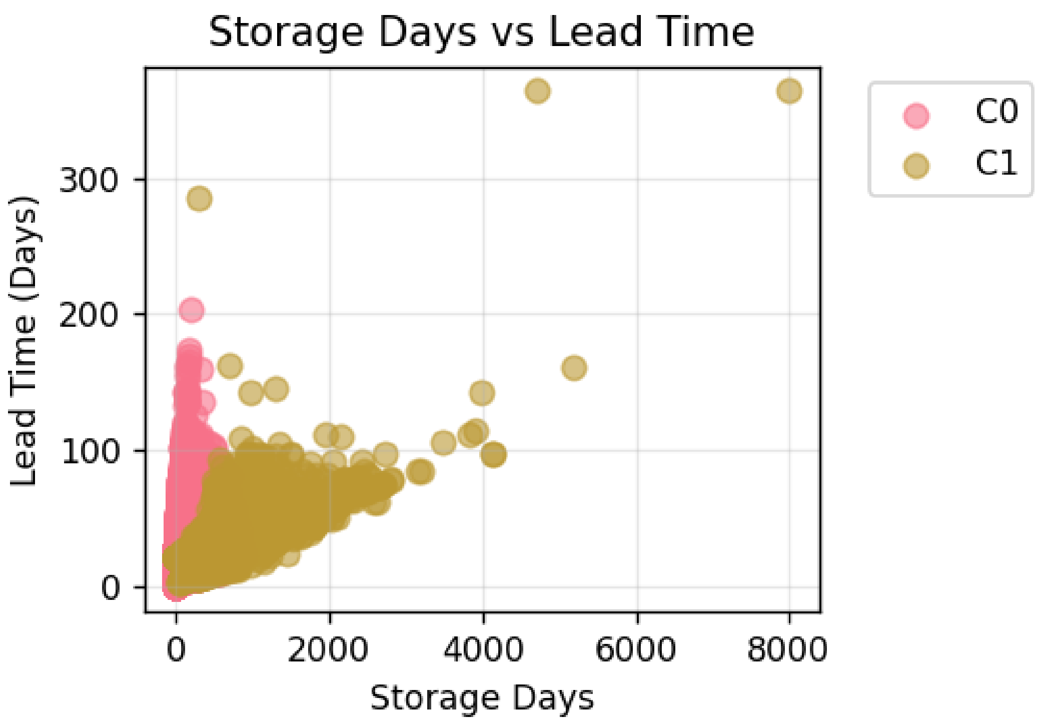


Figure 25 – Storage Days vs Lead Time by cluster, Python program output

This scatterplot compares the days a product remains in storage with the lead time (days from production to shipment). Each point is a shipment; colors indicate the K-Means clusters.

In the data a moderate positive relationship appears longer, in-stored shipments tend to have longer lead times. Cluster C0 (pink) is concentrated near the origin, with low storage days and short to medium lead times typical, well-flowing items. Cluster C1 (gold) ranges both to the right and upwards of that region indicating long lead times that are kept on hand for an extended duration. Some extreme cases emerge (very high storage or lead time) that must be audited: they may indicate exceptional flows, booking issues, or data anomalies. Operationally, the chart helps separate two dynamics.

- Near the origin, vertical spread at low storage suggests planning or network constraints (lead time longer even when storage is short).

- On the right, horizontal spread with high storage indicates aging stock waiting to be released, where FIFO and release rules matter more than transport time.

You may then use this view to define action thresholds that trigger priority release or escalation, or to flag lanes/products that most often sit in the upper-right area for focused interventions (booking earlier, consolidating differently, or revising safety stocks, for example).

6.6 Volume and density patterns

Volume measures also distinguish between the two groups. Average aging stock is approximately 148 in C0, and about 1,123 in C1, reflecting continued exposure in the minority cluster. Combined with weight, volume, and the volume-efficiency ratio, these are ways to interpret if constraints are weight-based, cube-driven, or driven by handling and temperature. Operationally, the profiling suggests two use cases. First, C1 could be considered a priority watchlist for specific actions when working on warehouse rotation or booking rules. Second, even within C0, DBSCAN's noise and micro-clusters can be deployed to flag small sets of records, for example by destination or temperature class, to warrant local follow-up.

Figure 26 shows the scatter of Weight versus Volume by cluster. The cloud for C0 is concentrated in the lower-left area (lower weights and volumes), while C1 extends to higher ranges on both axes. This is consistent with the correlation observed in the KPI heatmap (Weight–Volume ≈ 0.74): heavier batches tend to occupy more volume.

Interpreting the chart in terms of density (Weight/Volume) helps. Points along a steeper slope indicate higher packing density, while flatter regions indicate bulkier, low-density items. The spread of C1 suggests that a relevant subset combines high weight with large volume, i.e., shipments that can hit either payload limits or cube limits depending on the mix. In contrast, C0 concentrates more compact loads with lower absolute values.

Operational implications:

- Define density bands (e.g., low, medium, high) and use them in weekly planning to co-load compatible SKUs. High-density items are good candidates to fill residual payload in containers that are volume-limited; low-density items can fill cubic space in weight-limited containers.

- For flows with high weight and volume (typical of C1), check payload thresholds and cube capacity in booking rules to reduce under-utilization and rehandling.
- Consider temperature class when forming loads: reefer capacity and stowage constraints may shift the practical limit from volume- to weight-binding.
- Track density outliers (extremely high or low Weight/Volume) as they often drive sub-optimal utilization and higher cost per ton.

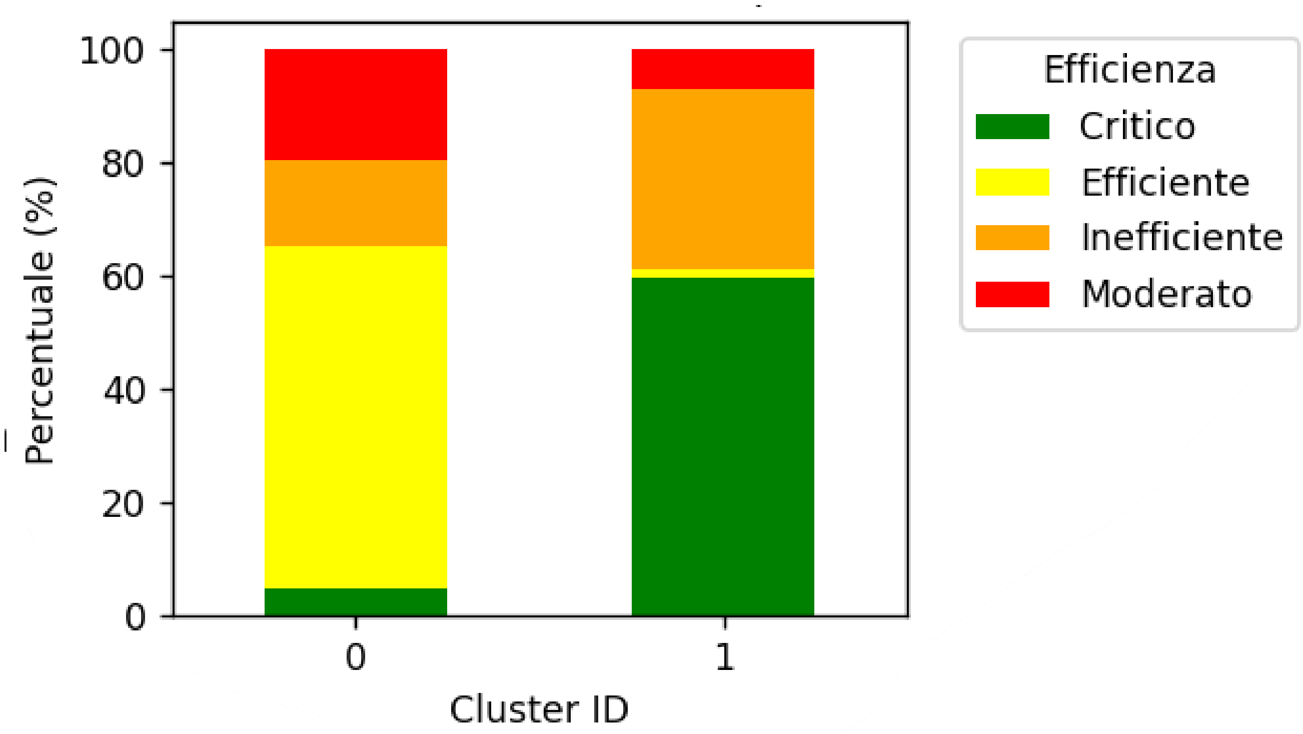


Figure 26 - Weight versus Volume by cluster, Python program output

The Figure 27 reports the composition of efficiency classes within each cluster.

The baseline cluster C0 shows a majority of Efficient and Moderate records, with a relatively small share of Inefficient and Critical cases, the minority cluster C1 presents the opposite profile. It concentrates most of the Critical and a large part of the Inefficient observations, while the Efficient share is much smaller.

This qualitative mix is coherent with the quantitative evidence seen earlier: C1 also exhibits higher storage days, longer lead time, and larger aging stock. Taken together with the KPI correlation matrix, these results point to a common driver.

The graph highlights that the storage accumulation is strongly associated with aging stock, and when combined with longer release times it produces the riskier patterns that characterize C1. The weight–volume chart suggests that some of these flows also operate near capacity constraints, which can amplify delays and stock ageing.

Operationally, the segmentation offers a simple triage. C0 can be managed with standard rules and periodic monitoring. C1 requires priority actions, for example faster rotation of slow movers, calendar alignment between warehouse release and carrier booking, and specific checks on volume and payload utilization for the affected lanes and temperature classes. This information will be used in the next chapter to quantify potential improvements in planning and to propose targeted interventions.

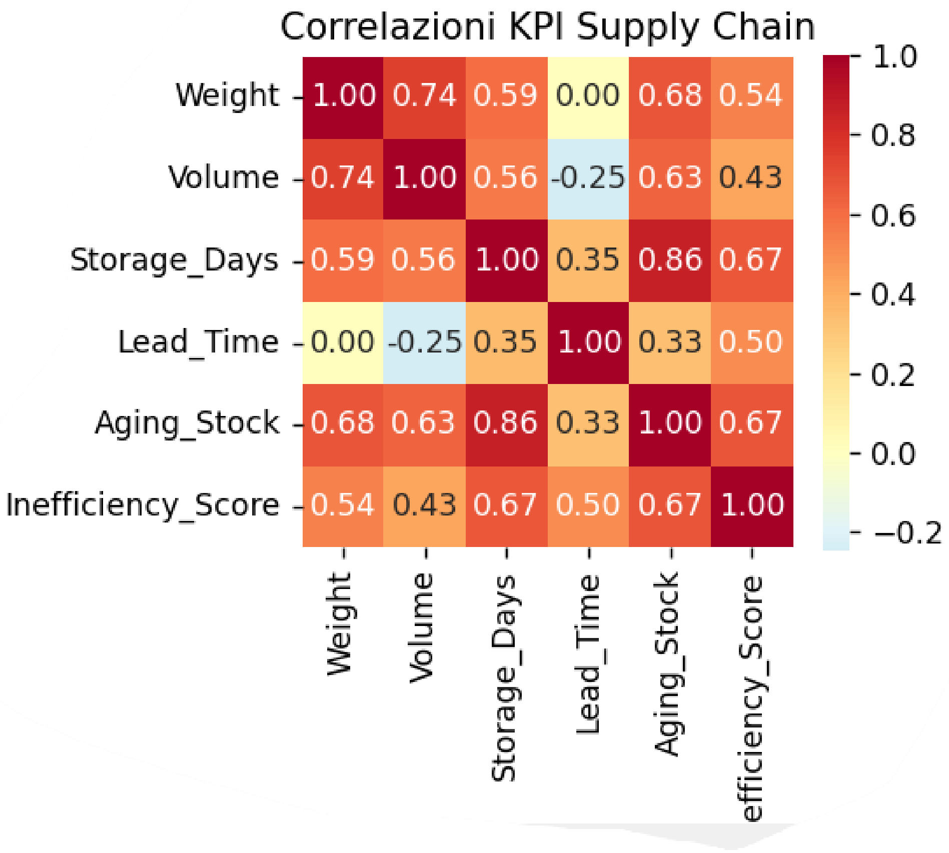


Figure 27 – Efficiency class composition by cluster, Python program output

Chapter 7 – Results

7.1 Results

Before any final points, it is useful to recap what was done and why. A dataset of about 78,000 rows and 29 columns was analyzed, tracking each product from factory release to shipment. Attention was placed on three simple logistics measures: days in storage (Storage Days), time from production to shipping (Lead-times), and the portion of inventory that was getting old (Aging Stock). To make the process clearer, a few additional indicators were defined: a storage pressure number (Storage Days × Weight), a turnover score (1/(Lead-times + 1)), and a combined Inefficiency Score that rises whenever storage, lead time, and aging are all high at the same time. The aim was descriptive, not predictive: the data was allowed to show its own shape and highlight where time and risk accumulate. In the absence of any label that says "this line is a problem," unsupervised clustering was applied: K-Means to obtain a big-picture map and DBSCAN to find small dense pockets and outliers. PCA was used only to clean up the charts; it did not change the underlying numbers. In short, the split reveals a large "normal" cluster and a smaller "hot spot" where storage, lead time, and aging concentrate; this is the natural target for quick wins.

When lines were grouped by the Inefficiency Score (High Storage + Long Lead Time + High Aging), the pattern became clear: 57.5% Efficient, 19.0% Moderate, 16.1% Inefficient, and 7.5% Critical.

At the system level, the process clearly splits into two patterns. The K-Means routine uses K = 2 and reports a silhouette score of 0.504, which signals a clear split. Cluster 0 (C0) covers 95.1% of the rows (74,404 entries) and matches the baseline: Storage Days ≈ 105.1, Lead-times ≈ 18.2, Aging Stock ≈ 148.0, Inefficiency Score ≈ 0.64. Cluster 1 (C1) makes up 4.9% (3,860 rows) yet concentrates most of the time and risk: Storage Days ≈ 659.7, Lead-times ≈ 32.3, Aging Stock ≈ 1,123.4, Inefficiency Score ≈ 2.49.

All values reported for the two clusters come directly from the data after K-Means assigned a label to every line. Cluster shares were obtained by counting lines per label and dividing by the total number of lines (N = 78,264):

C0 share = 74,404 / 78,264 ≈ 95.1%; C1 share = 3,860 / 78,264 ≈ 4.9%.

Within each cluster, the key KPIs were computed as ordinary arithmetic means over the rows belonging to that cluster. If $C_j$has $n_j$lines, the mean of KPI $X$is

$$\mathrm{Avg}_X(C_j) = \frac{1}{n_j} \sum_{i \in C_j} X_i.$$

Applied to the dataset, this yields:

- Cluster 0 (C0) : Storage Days(C0) ≈ 105.1 (average of Storage Days over its 74,404 lines); Lead-times(C0) ≈ 18.2; Aging Stock(C0) ≈ 148.0.
- Cluster 1 (C1) : Storage Days(C1) ≈ 659.7 (average over its 3,860 lines); Lead-times(C1) ≈ 32.3; Aging Stock(C1) ≈ 1123.4.

Reporting the Inefficiency Score, described in chapter 5.6, a simple 0–3 sum of three binary flags defined at the line level using the 75th percentile (Q75) from the full dataset:

- High_Storage = 1 if Storage_Days > Q75(Storage Days), else 0;
- Long_Lead_Time = 1 if Lead_Time > Q75(Lead-times), else 0;
- High Aging = 1 if Aging_Stock > Q75(Aging Stock), else 0.

For each cluster, the reported number is the average of the 0–3 score across its lines:

$$\mathrm{AvgScore}(C_j) = \frac{1}{n_j} \sum_{i \in C_j} \mathrm{Inefficiency\_Score}_i.$$

This gives ≈ 0.64 for C0 and ≈ 2.49 for C1. The higher average in C1 reflects that many C1 lines exceed the 75th-percentile thresholds on one or more dimensions (storage, lead time, aging). All "≈" symbols indicate standard rounding (typically one decimal place) for readability.

Running DBSCAN (eps = 0.5) in parallel finds about 713 dense pockets and marks 14.4% of the observations as outliers, highlighting local quirks a global split might miss. Together, these signals point to a broad "basin" and a tighter "hot spot" where delays and aging pile up. All of these values come from the feature matrix used by the routine. That matrix includes the KPIs (Storage Days, Lead-times, Aging Stock, Weight, Volume, etc.) along with engineered fields:

- Storage_Cost_Index = Storage Days × Weight,
- Stock_Rotation_Efficiency = 1/ (Lead-times + 1),
- Expiration Risk = Aging Stock/ (Storage Days + 1),
- Volume Efficiency = Weight/ (Volume + 0.001).

The categorical attributes (Product Type, Division, Warehouse, Destination Country, Strategy, Temperature) were turned into columns via label encoding. Before clustering, missing values in the feature matrix were filled with each column's median; then StandardScaler was used to convert all features into z-scores.

K-Means was fit over k = 2...9. For each k we computed the silhouette score on the data matrix; the silhouette for any observation is (b − a)/max(a, b) where a is the distance to points in the same cluster and b the average distance to points in the nearest other cluster. The highest average silhouette appeared at k = 2 (0.504), so we chose K = 2. After K-Means assigned labels, the counts were n_0 = 74,404 and n_1 = 3,860 out of a total N = 78,264 lines. This gives C0 ≈ 95.1% (74,404/78,264) and C1 ≈ 4.9% (3,860/78,264). Within-cluster means were taken for each KPI. So, C0 has Storage Days ≈ 105.1, Lead-times ≈ 18.2, Aging Stock ≈ 148.0; by contrast, C1 shows higher values — Storage Days ≈ 659.7, Lead-times ≈ 32.3, Aging Stock ≈ 1,123.4.

We built three binary flags from dataset quartiles: High Storage (1 if Storage Days > Q_{75}), Long_Lead_Time (1 if Lead-times > Q_{75}), High Aging (1 if Aging Stock > Q_{75}). The Inefficiency Score is the sum of these flags (0–3). The cluster averages — ≈ 0.64 for C0 and ≈ 2.49 for C1 — are the means of this score within each cluster.

DBSCAN was run with min-samples = 5, testing ε in {0.5, 1.0, 1.5, 2.0}. The first ε that gave structure without too much noise was 0.5. In DBSCAN, a label of −1 means outlier; any other label means a cluster. At ε = 0.5 the run found 713 clusters and flagged 11,287 points as outliers (11,287/78,264 ≈ 14.4%).

To link clusters, analyzing how variables move together, we used Pearson correlations. Pearson returns a coefficient r (which expresses the correlation between two variables) from −1 to +1 for linear association:

- values near +1 mean the two variables rise together
- values near 0 mean little linear link
- values near −1 mean they move opposite

Correlations come from the line-level table in Chapter 6 (one row per product line) and use: Storage Days, Aging Stock, Lead-times, Inefficiency Score, Weight, Volume. The analysis is pairwise: for any variable pair, only rows with both measures are kept. This matters because Weight and Volume have many missing values, so any correlation with them uses a smaller sample. We didn't add any extra scaling or fill missing values at this point, **because** Pearson already standardizes each pair. We left extreme values as they are, without capping or trimming. This keeps the result easy to read, though big or unusual numbers can still affect the correlations.

All the correlations between the variables suggests a different next step:

Storage Days ↔ Aging Stock: This checks if lines that stay longer in the warehouse build up aging inventory. A strong positive correlation means they rise together, a near zero correlation means no clear link, and a negative coefficient would mean more storage, less aging. In short, a strong positive link suggests that reducing storage should cut expiry risk.

Storage Days ↔ Inefficiency Score: Analyze whether longer storage is associated with higher inefficiency scores. A positive coefficient means longer storage usually adds to the score, which fits how the metric is built. A near zero means storage by itself doesn't explain the score and a negative reading would be odd here. A strong positive link supports prioritizing actions that cut storage days.

Aging Stock ↔ Inefficiency Score: Checks whether older inventory gets higher scores. A positive correlation supports aging as a driver, a near zero means age barely relates and negative correlation signals a data/scoring issue. A strong positive link backs tighter FEFO (first expired, first out)and faster release of fast-aging lines.

Lead-times ↔ Inefficiency Score: This score sees whether the longer time from production to shipment relates to inefficiency. A positive coefficient means longer lead times usually come with higher scores, though other drivers (like storage days) matter, too. Near zero means lead time barely matters. A positive link highlights better hand-offs and fewer stalls between production, booking, and shipping.

Lead-times ↔ Storage Days: This question asks whether a longer lead time comes with longer storage. A positive reading means a slow release goes with more storage; near zero means they're mostly independent; a negative reading would be unusual. A positive link suggests aligning production with warehouse/booking windows to avoid one delay amplifying the other.

Weight ↔ Volume. Check whether heavier products are also bigger. In practice, a strong positive correlation is common: as items get heavier they take up more space. When the correlation is near zero, weight and volume move separately. A tight link raises the risk of hitting weight or cube limits early, reducing co-loading and lengthening waits — hence the push for a balanced trade-off between weight and cube.

Before looking at the data, it's helpful to explain how we calculate the percentages and averages in figures 28 and 29. In both figures, "100%" means the total for the dataset. In Figure 28, the baseline is the sum of all storage-days (one storage-day = one line kept for one day). For each cluster, we multiply the line count by that cluster's average Storage_Days; then we divide each result by the sum across both clusters and show it as a percentage. Using the numbers: C0 = 74,404 × 105.1 and C1 = 3,860 × 659.7; together they add up to the whole (100%). As shares of that total, C0 accounts for about 75.5%, while C1 covers about 24.5%. Figure 29 uses the same steps for Aging Stock: line count per cluster × average aging in that cluster, each divided by the combined aging of both clusters, which gives ≈ 71.7% for C0 and ≈ 28.3% for C1. In short, C1 doesn't show up often, but its high averages make it account for a large share of the total.

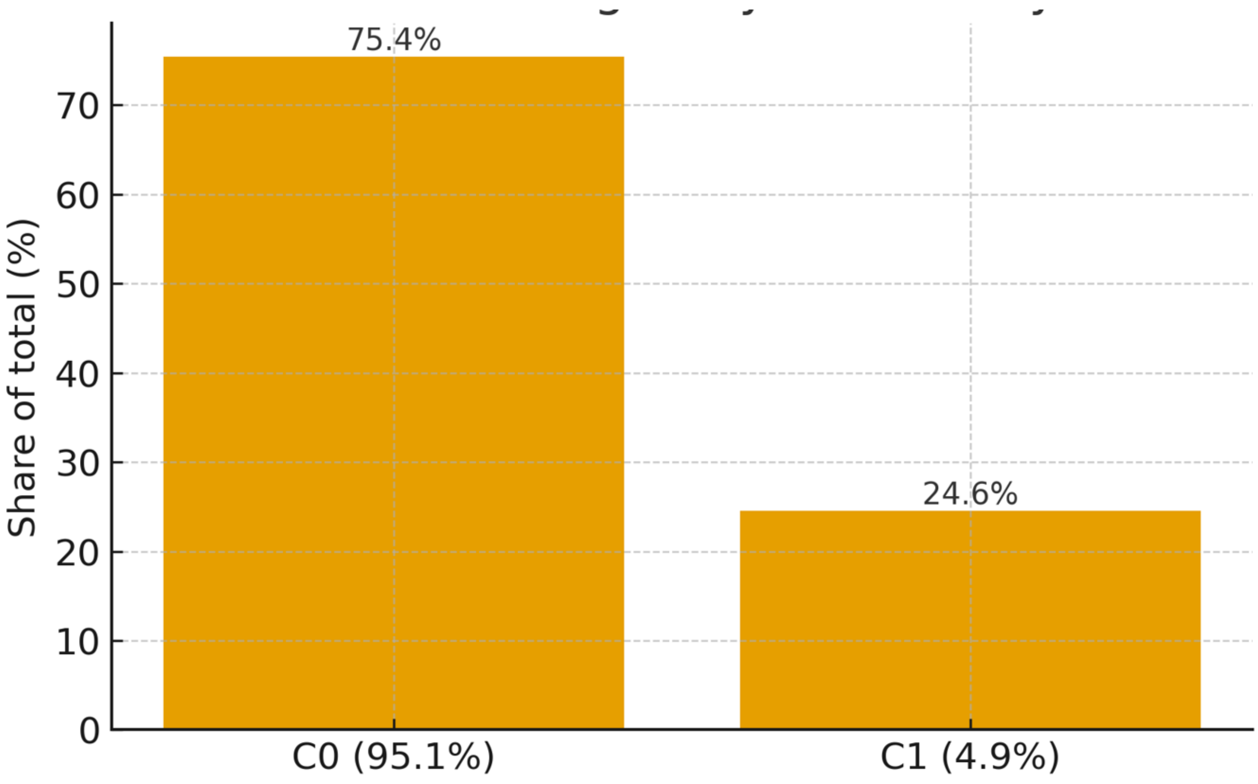


Figura 28 – Share of the total storage days by cluster, Python program output

This chart above splits 100% of the storage time in the whole system between Cluster 0 and Cluster 1. Here, “100%” means the sum of all storage days across every line in the dataset, where one storage day is one line staying one day in the warehouse. The percentage for each cluster is calculated as (number of lines in the cluster × its average storage days) divided by (total storage days of all lines). With our data: C0 has 74,404 lines × 105.1 days, C1 has 3,860 lines × 659.7 days; adding these two gives the total (the 100%). From this we get about 75.5% for C0 and 24.5% for C1. The meaning is clear: even if C1 is only 4.9% of all lines, it creates about a quarter of the storage time. So, reducing storage days in C1 changes the global average much faster than the same effort on C0.

The same logic can be extended to expiry risk. By scaling each row with the Aging Stock we uncover each cluster’s slice of the aging stock.

Figura B

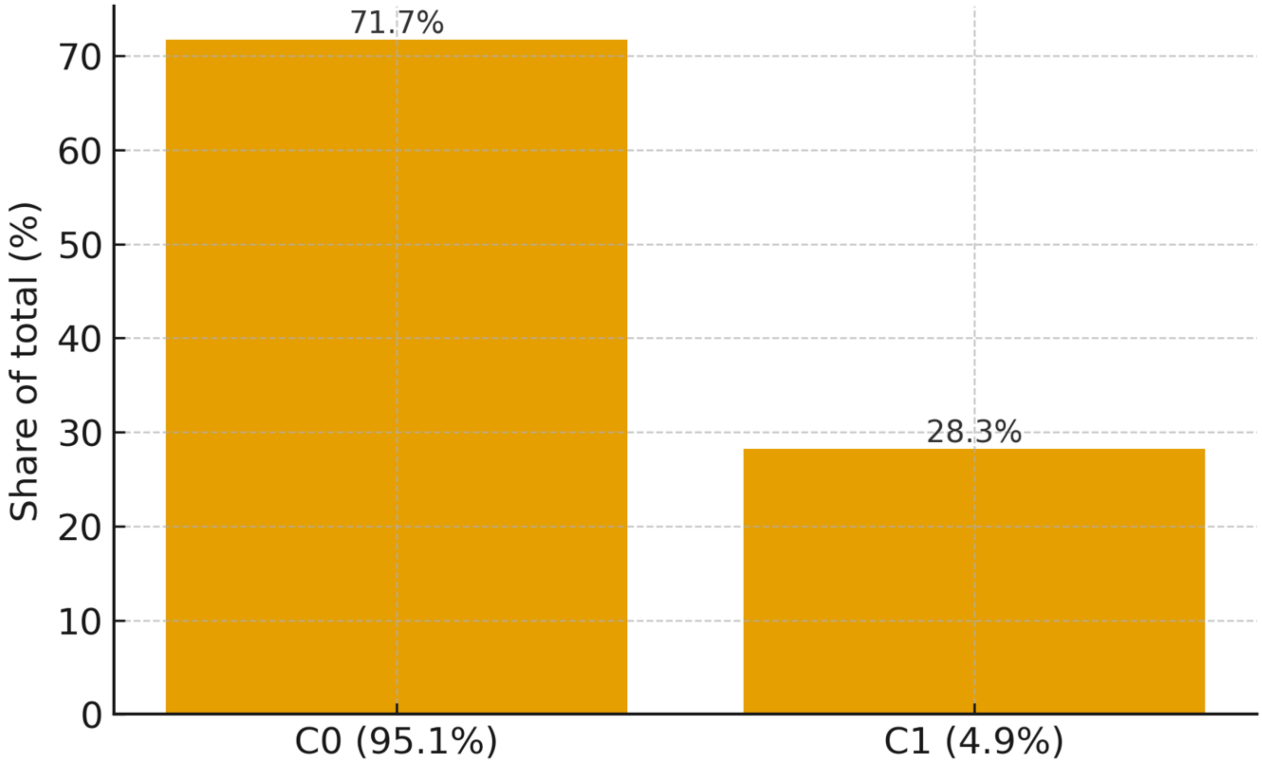


Figure 29 - Share of total aging stock by cluster, Python program output

Figure B above, reports the aging stock between the two clusters. Here, “100%” means all aging stock summed across every line. The percentage for each cluster is (number of lines in the cluster × its average aging) divided by the total aging of all lines. Using our numbers, C0 contributes 74,404 × 148.0, C1 contributes 3,860 × 1123.4; their sum is the total (the 100%). The shares are about 71.7% for C0 and 28.3% for C1. The interpretation is straightforward: C1 is small but carries almost one third of the expiry risk. In daily operations this supports strict FEFO and earlier release for C1 items and a booking plan that avoids extra waiting once these pallets are ready.

To help with planning, we also built a scenario view that shows how the storage changes whenever a target for C1 is chosen.

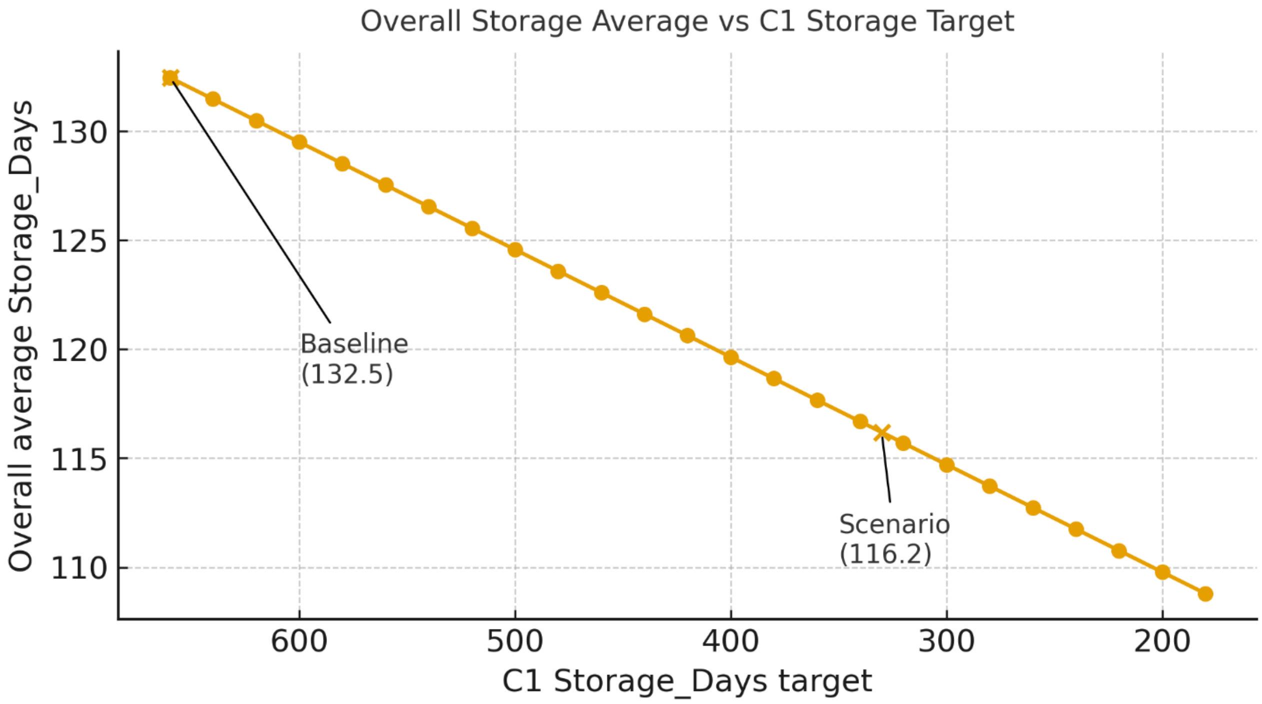


Figure 30 - Overall storage average vs C1 storage target

The overall average is calculated in Figure 30 (today vs. scenario). The network's overall storage is a weighted average of the two cluster averages, with the cluster shares as weights. Let p0 ≈ 0.951 and p1 ≈ 0.049 be the shares of clusters C0 and C1 and let μ0 = 105.1 and μ1 = 659.7 be their storage averages.

Baseline (today): Overall today = p0×μ0 + p1×μ1. Plugging in p0 = 0.951, μ0 = 105.1, p1 = 0.049, μ1 = 659.7 gives about 132.3 days.

Scenario: If C0 stays fixed and we set a new target T for C1, the formula is: Overall(T) = 0.951×105.1 + 0.049×T.

When T = 330, the result is about 116.1 days. That's a drop of about 16.2 days (≈ 12.2%) from today. Lowering the C1 average pulls down the overall average, making "reduce stock" a clear number for planning.

The plot shows clearly how the overall average storage changes when you set a target for C1's storage. The overall average is just a mix of the two clusters: 95.1% × C0 average (fixed at 105.1) + 4.9%×C1 target. Right now, with C1 around 659.7 days, the combined average is about 132.3 days; lowering C1 to 330 days would reduce it to 116.1 days. Because the x-axis goes backward, sliding right means choosing an ambitious C1 goal as a consequence we can see that the line drops slowly as a lower C1 pulls the total down. In practice, a manager can set a C1 target and instantly see the expected overall storage. For example, cutting it from 660 days to about 330 days means a 12% drop overall.

To sum up, the figure 31 below reports a short comparison summary of today's numbers and the scenario view, showing storage, lead time, and aging in one view.

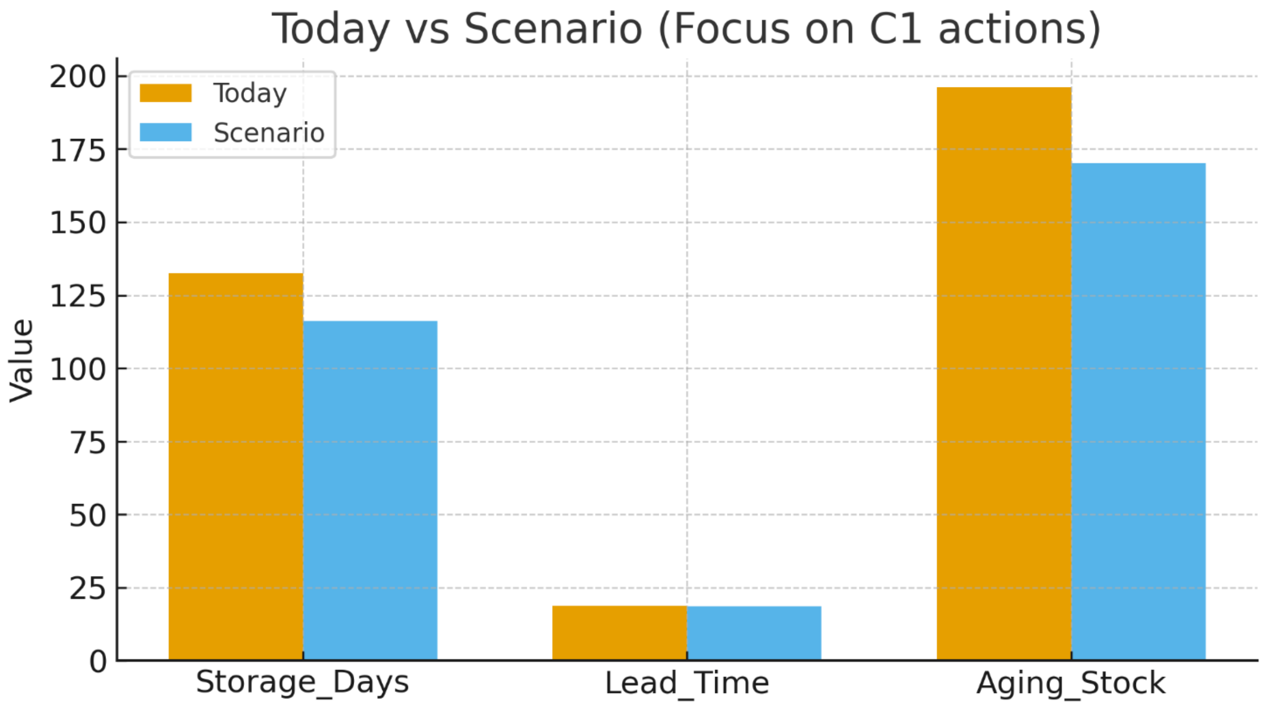


Figure 31 — Today vs scenario (focus on C1 actions)

The chart shows the average for three main measures (Storage Days, Lead-times, and Aging Stock) against the overall average, in a simple scenario that only changes C1. Each overall figure comes from a mix: 95.1% × C0 average + 4.9% × C1 value.

In the "Today" view the actual C1 numbers are used (storage = 659.7, lead time = 32.3, aging 1123.4). In the "Scenario" C0 stays the same while the C1 goals are set to storage = 330, lead time= 25 and aging = 600

The chart's bars show storage dropping from 132.3 to about 116.1, aging falling from about 195.8 to around 170.1, and lead time going down a little from about 18.9 to 18.5. In short, focusing on C1 brings overall gains, especially for storage and expiration risk.

7.2 Analysis overview

Seen from a point of view, combining K-Means with DBSCAN gives two views on the same process and together they tell one clear story. K-Means shows the idea: after scaling the data it splits the data into a large "normal" group and a small hotspot where storage days, lead time and aging all go up together. That small hotspot isn't noise; it's where time and risk pile up. DBSCAN, which finds dense groups without setting k first, works like a close-up on the data. It finds many small groups and a few outliers that a big split would miss, and those pockets often match details — like a route, a temperature group, a pack type or a repeat booking pattern that keeps adding waiting time. Together, the two methods aren't rivals. They are two layers of the same picture: a big map that shows where to look first and a zoomed-in view that shows exactly where a rule or an allocation should change.

This two-layer view fits basic operations logic. The longer an item stays in the system the more inventory grows. That extra stock raises the chance of aging or waste. In simple terms the chain reads "time → inventory → risk." Chapter 7.1's correlation data show it in numbers: storage days, aging and the overall score all go up together while lead time adds a bit. Variation also matters as regular flows and use buffers, whereas irregular flows need extra buffers. The small cluster is where things vary a lot and work together less, for example, special items, tighter temperature rules, or extra paperwork and booking rules that stop us from loading shipments together. Then comes capacity limits: when loads hit weight or space limits the chance to combine loads goes down, so an early delay can become a longer later wait. These points explain why a small share of lines can cause a lot of total storage time and aging and data don't prove cause. The same pattern, across methods and metrics, suggests a lesson: first fix the small, tight group that causes delay and the network will get better faster.

What makes the result useful is that the same pattern can be turned into steps. The K-Means map highlights top actions and sets number targets. For example, cutting in half the cluster's storage, from roughly 660 days to 330 days, would lower the overall average by about 12%. DBSCAN focuses on the spots where a reduction is most likely on tight booking lanes, temperature classes that stress cold capacity, or mixes where weight and space clash. In this sense, the analytics don't replace judgment; they support it, giving a starting point for pilots and a simple way to measure impact.

Chapter 8 – Conclusions

This thesis used machine learning clustering to improve supply chain performance in the Ferrero Group case study. It began with a review of basic logistics and supply chain ideas, which showed the importance of transport, storage, and inventory management in the food industry. The practical part looked at three main performance indicators: days in warehouse, total lead time from production to shipment, and risk of aging stock. After cleaning and preparing the dataset, two clustering methods were used: K-Means and DBSCAN. The results show a clear split: 95% of products behave normally, with a good storage and lead-time values, and 5% behave very differently (very long storage times, longer lead times, and a high expiry risk). Based on these findings, K-Means chose k = 2 with a silhouette score of 0.504, separating a large main cluster (95%) from a tight hotspot (5%). Products in the hotspot had about 660 average days in storage, about 32 days from production to shipment, and 1123 units of aging exposure, showing a focused pocket of inefficiency. DBSCAN, used as a density-based check, confirmed this hotspot and found small high-density pockets and outliers linked to lanes, temperature classes, or pack types. It's shown that cutting the average storage time in the hotspot in half (≈ 660→330 days) would lower the overall network average by about 12%, showing the big payoff of focused actions.

Unsupervised clustering works well to spot operational problems in a food logistics network and rank which fixes to do first. The performance does not drop because of many small issues everywhere, but it comes from a small set of SKUs with extreme behaviour. From a methods view, this study shows that a simple, clear pipeline, with median imputation for missing values, standardizing numbers, label-encoding categories, K-Means for overall structure, DBSCAN for local density, and cluster profiling, can give useful guidance with low model complexity. The two methods are complementary: K-Means gives a stable global split for monitoring and KPI tracking, and DBSCAN helps find local risk pockets linked to real-world conditions (e.g., certain lanes or temperature settings).

8.1 Limitations of the analysis

The dataset is very large but not perfect. For example, Weight and Volume have many missing values, and for this before the clustering analysis, they were filled with the median imputation, and the data were scaled. The Median imputation (described in paragraph 5.3 Data cleaning) is when missing values are replaced with the middle value, making the data more stable but reducing spread (in practice: it pulls extreme values toward the center, so differences between lines can look smaller). Standardization means putting all variables on the same scale (average 0, spread 1), so no single variable overpowers the others just because it has bigger units.

Categorical fields (product type, division, warehouse, destination, strategy, temperature) were turned into numbers so the model could use them. Label encoding turns categories (like countries) into numbers (e.g., A→0, B→1, C→2) but it creates a fake order (the model may treat “2” as larger than “1” even if the categories have no real ranking). Other encodings (like target or frequency encoding) or a distance that handles numbers and categories could handle categories better and might change the exact cluster limits.

Model choices also have limits. For example K-Means works better when clusters are round after scaling and when each variable has a similar spread (it relies on straight-line distance), and it can be easily affected by outliers. DBSCAN depends a lot on its two settings: *eps* (how close points must be to count as neighbors) and *min_samples* (how many neighbors are needed to call an area “dense”). Changing these values can split one area into small groups or merge separate groups. Pearson’s *r* measures only linear links and if the true relation is not linear (for example, a cutoff where problems jump after a certain day), *r* may look weak even if the pattern is real.

Costs are not included in the dataset and in the analysis, so gains are measured in days saved and fewer aging units, not in euros. Adding storage costs per day, power bills, detention/demurrage fees, and penalties, the same scenarios can become a cost view and then actions can be ranked by effect and payback time.

Time patterns are not analyzed in depth , in fact, the analysis uses snapshots and totals, so seasonality, booking calendars, and route-specific quirks show up only indirectly unless they are added as variables we add. An over-time view (for example, weekly C1 share by lane and temperature) would separate one-off spikes from recurring patterns.

The code is flexible but not ready to use with every kind of dataset, as the one used was built and tuned step by step. About one month went into cleaning inputs, matching columns, defining KPIs, setting thresholds, tuning clustering, and making plots stable. The script now runs from start to finish on this dataset, but if input columns change or fields are added/removed, parts of the mapping, made KPIs, and charts will need updates. A settings file (to hold column names, thresholds, and model parameters) and unit tests would speed up and reduce risk for future changes.

8.2 Suggestions for future research

At this point the work moves to linking time to money, so choices are guided not just by the count of days or the aging of units but by the euro cost attached to them. The storage cases previously outlined can be rebuilt by adding a storage fee, including the power use of cold-rooms, adding detention and demurrage charges, and adding any service penalties. When those settings are in the model, the "today vs target" curves turn into predictions, allowing actions to be ranked by their expected impact and payback time. Another path is to add time to the features. A weekly view of the small cluster share by lane and temperature separates peaks, booking patterns, and ongoing bottlenecks from one-off spikes, making it easier to find where to intervene and when to plan extra capacity.

Another point could be adding shipment-level weight and space use into the features, along with compatibility tags, like destination and temperature, the system could reduce big bottlenecks that cause long waits. This approach might also show which combinations create space and which rules allow co-loading safely. The fourth improvement needs a pipeline that's more flexible and reliable. Moving column mappings, thresholds (like the 75th percentiles), and model parameters (K range, DBSCAN eps and min_samples) to a configuration file, and adding a few unit tests, would make adaptation to new datasets faster and safer.

It makes sense to try algorithms that can handle numbers and categories together and clusters that are not round. Our dataset mixes numeric KPIs (days in storage, lead time, aging, weight, volume) with context (product type, warehouse, destination, strategy, temperature). HDBSCAN is a good fit because it adapts what "dense" means to the data instead of using one fixed rule [36]. In practice, this helps us see if the small hot spot is a real, stable group, and if there are other tiny groups nearby that a simple split would miss. Instead of giving each line a fixed label, Gaussian Mixture Models give a probability that helps when the border between the baseline and the hot spot is fuzzy. Agglomerative clustering is another option because it can use a distance that works with both numbers and categories, so it can find structure at more than one level. It may show two clusters that match K-Means, and inside the hot spot, it can find subgroups linked to lanes, temperature classes, or pack types. These methods are not meant to replace the current baseline as they are a stability check and add another view of the same pattern.

Bibliography

[1] Introduction to distribution logistics, Brandimarte & Zotteri, 2004.

[2] R. Bowersox, J. Closs and M. B. Cooper, “Supply Chain Logistics Management”

[3] Defining supply chain management, Mentzer, 2001.

[4] Integrating the supply chain, Stevens (1989)

[5] A taxonomy of green supply chain management capability, Shang, Lu, and Li (2010)

[6] Green Supply Chain Management Strategies, Lean Six Sigma Glossary term, https://sixsigmadsi.com/glossary/green-supply-chain-management

[7] Green supply chain management: Pressures, practices, and performance - An integrative literature review, Virendra Balon, 2019.

[8] Ceresio Investors – Food Industry Monitor 2024

[9] Consilium comunicazione Food Industry Monitor 2024 Summary, Published June 27, 2024.

[10] Croom, S., Cox, A., Chicksand, D., & Yang, T. (2007). The proactive alignment of sourcing with marketing and branding strategies: a food service case. Supply Chain Management: An International Journal, 12(5), 321-333.

[11] Thron, T., Nagy, G., & Wassan, N. (2007). Evaluating alternative supply chain structures for perishable products. The International Journal of Logistics Management, 18(3), 364-384.

[12] Fredriksson, A., & Liljestrand, K. (2015). Capturing food logistics: a literature review and research agenda. International Journal of Logistics Research and Applications, 18(1), 16-34.

[13] Theodoras Varzakas, T. (2006). Handbook of Food Processing: Food Preservation (chapter on storage temperature and logistics). CRC Press.

[14] Fredriksson, A., & Liljestrand, K. (2015). Capturing food logistics: a literature review and research agenda. International Journal of Logistics Research and Applications, 18(1), 16-34.

[15] Navickas, V., Baskutis, S., Gruzauskas, V., & Kabasinskas, A. (2016). Warehouses consolidation in the logistic clusters: Food industry’s case. Polish Journal of Management Studies, 14(1), 174-183.

[16] IEA (2024) World Energy Balances and Renewables Information

[17] Euroviews - Electric cars are still cheaper to run than petrol and diesel. https://www.euronews.com/2023/07/24/electric-cars-are-still-cheaper-to-run-than-petrol-and-diesel

[18] ISPRA - National Greenhouse Gas Emissions Inventory

[19] Euostatistics - Freight transport statistics, Modal split of freight transport in the EU

[20] The Geography of Transport Systems, Modal Share of Freight Transportation, Selected Countries

[21] Introduction to Materials Management, Arnold, Chapman and Clive (2007)

[22] Sea freight then and now - the history of maritime forwarding development, Real Logistics

[23] A container ship is docked, and its cranes are loading containers on board at the quay. https://commons.wikimedia.org/wiki/File%3AContainer_Ship.jpg

[24] Analysis, modeling, and assessing performances of supply chains served by long-distance freight transport corridors, Bart Wiegmans

[25] The history of containers, MC Containers, https://mccontainers.com/blog/the-history-of-containers/

[26] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & others. (1996).. In kdd (Vol. 96, pp. 226–231).

[27] Clustering Like a Pro: A Beginner’s Guide to DBSCAN, https://medium.com/@sachinsoni600517/clustering-like-a-pro-a-beginners-guide-to-dbscan-6c8274c362c4

[28] "Lloyd, Stuart P. \"Least squares quantization in PCM.\" *Information Theory*, IEEE Transactions on 28.2 (1982): 129-137".

[29] Introduction to K-Means Clustering, https://www.bombaysoftwares.com/blog/introduction-to-k-means-clustering

[30] DBSCAN - Density Based Spatial Clustering of Applications with Noise, https://ubc-library-rc.github.io/ml-classification-clustering/content/dbscan.html

[31] FERRERO GROUP REPORTS CONSOLIDATED FINANCIAL STATEMENTS FOR THE 2023/2024 FINANCIAL YEAR, https://www.ferrero.com/int/en/news-stories/news/ferrero-group-reports-consolidated-financial-statements-for-the-2023-2024-financial-year

[32] *Ferrero global sites with published receiving hours on the Suppliers portal.* Source: Ferrero Suppliers, Delivery and Tracking. https://www.ferrerosuppliers.com/en/delivery-and-tracking

[33] Ferrero Group announces emissions reduction, https://sustainabilityonline.net/news/ferrero-group-announces-emissions-reduction-renewable-energy-achievements

[34] Briano, E., Caballini, C., Giribone, P., & Revetria, R. (2010). Using a system dynamics approach for designing and simulation of short life-cycle products supply chain. Proceedings of the 4th WSEAS International Conference on System Science and Simulation in Engineering (ICOSSSE'10), Genova, Italy. ISSN 1970-5117.

[35] Fisher M.L. 1997 What is the right supply chain for your product? Harvard Business Review.

[36] HDbscan: Hierarchical density based clustering, Leland McInnes, John Healy, https://joss.theoj.org/papers/10.21105/joss.00205?

Figures and tables lists

Appendix

Python program:

```
import argparse
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.cluster import KMeans, DBSCAN
from sklearn.metrics import silhouette_score
from sklearn.decomposition import PCA
import warnings
from datetime import datetime
import os
from sklearn.impute import SimpleImputer


warnings.filterwarnings('ignore')

# Setup matplotlib
plt.style.use('default')
sns.set_palette('husl')


def load_and_prepare_data(file_path):
  """Carica e prepara i dati per l'analisi di clustering senza gestione missing/duplicati"""
  print('='*70)
  print('PHASE 1 - CARICAMENTO E PREPARAZIONE DATI')
  print('='*70)

  if not os.path.exists(file_path):
    print(f"File non trovato: {file_path}")
    return None

  # Caricamento robusto del CSV
  try:
    df = pd.read_csv(file_path, low_memory=False, encoding='utf-8')
  except:
    try:
      df = pd.read_csv(file_path, low_memory=False, encoding='utf-8', sep=';')
    except:
      try:
        df = pd.read_csv(file_path, low_memory=False, encoding='iso-8859-1')
      except:
        df = pd.read_csv(file_path, low_memory=False, encoding='cp1252')

  print(f"Dataset caricato: {df.shape[0]:,} righe, {df.shape[1]} colonne")

  # Normalizza nomi colonne
  df.columns = [str(c).strip() for c in df.columns]
```

```
    # Mappa colonne esistenti a nomi standardizzati
    column_mapping = {
        # Date
        'Calcolo Data PROD': 'Data_Produzione',
        'Data Spedizione Disp.': 'Data_Spedizione',
        'Calcolo Data MATU': 'Data_Maturazione',

        # Metriche numeriche chiave
        'GG da PRODUZIONE': 'Days_Production',
        'GG da MATURAZIONE': 'Days_Maturation',
        'Actual Storage': 'Storage_Days',
        'Aging Stock': 'Aging_Stock',
        'Peso': 'Weight',
        'Volume': 'Volume',

        # Categoriche per clustering
        'Testo breve materiale': 'Product_Type',
        'Divisione': 'Division',
        'Magazzino': 'Warehouse',
        'Nazione di Destinazione': 'Destination_Country',
        'Push/Pull': 'Strategy',
        'Temperatura': 'Temperature'
    }

    # Rinomina colonne disponibili
    for old_name, new_name in column_mapping.items():
        if old_name in df.columns:
            df[new_name] = df[old_name]
            print(f"Mappata colonna: {old_name} -> {new_name}")

    # Conversioni specifiche
    # Date
    date_cols = ['Data_Produzione', 'Data_Spedizione', 'Data_Maturazione']
    for col in date_cols:
        if col in df.columns:
            df[col] = pd.to_datetime(df[col], dayfirst=True, errors='coerce')

    # Numeriche
    numeric_cols = ['Days_Production', 'Days_Maturation', 'Storage_Days',
                    'Aging_Stock', 'Weight', 'Volume']
    for col in numeric_cols:
        if col in df.columns:
            df[col] = pd.to_numeric(df[col], errors='coerce')

    # Calcola lead time se possibile
    if 'Data_Produzione' in df.columns and 'Data_Spedizione' in df.columns:
        df['Lead_Time'] = (df['Data_Spedizione'] - df['Data_Produzione']).dt.days
        print("Calcolato Lead_Time (giorni prod->spedizione)")

    print(f"Preparazione completata - colonne standardizzate disponibili")
    return df


def calculate_supply_chain_kpis(df):
```

```
"""Calcola KPI specifici """
print('\n' + '='*70)
print('PHASE 2 - CALCOLO KPI SUPPLY CHAIN')
print('='*70)

# KPI di costo storage
if 'Storage_Days' in df.columns and 'Weight' in df.columns:
    df['Storage_Cost_Index'] = df['Storage_Days'] * df['Weight']
    print("Calcolato Storage_Cost_Index = Storage_Days * Weight")

# Efficienza rotazione stock
if 'Lead_Time' in df.columns:
    df['Stock_Rotation_Efficiency'] = 1 / (df['Lead_Time'] + 1)
    print("Calcolato Stock_Rotation_Efficiency = 1/(Lead_Time + 1)")

# Risk di scadenza
if 'Aging_Stock' in df.columns and 'Storage_Days' in df.columns:
    df['Expiration_Risk'] = df['Aging_Stock'] / (df['Storage_Days'] + 1)
    df['Expiration_Risk'] = df['Expiration_Risk'].clip(0, 10)  # Cap outliers
    print("Calcolato Expiration_Risk = Aging_Stock / (Storage_Days + 1)")

# Efficienza volumetrica
if 'Weight' in df.columns and 'Volume' in df.columns:
    df['Volume_Efficiency'] = df['Weight'] / (df['Volume'] + 0.001)  # Evita div by zero
    print("Calcolato Volume_Efficiency = Weight / Volume")

# Score composito di inefficienza
inefficiency_components = []

# Componente 1: Storage elevato
if 'Storage_Days' in df.columns:
    storage_threshold = df['Storage_Days'].quantile(0.75)
    df['High_Storage'] = (df['Storage_Days'] > storage_threshold).astype(int)
    inefficiency_components.append('High_Storage')

# Componente 2: Lead time lungo
if 'Lead_Time' in df.columns:
    leadtime_threshold = df['Lead_Time'].quantile(0.75)
    df['Long_Lead_Time'] = (df['Lead_Time'] > leadtime_threshold).astype(int)
    inefficiency_components.append('Long_Lead_Time')

# Componente 3: Alto aging stock
if 'Aging_Stock' in df.columns:
    aging_threshold = df['Aging_Stock'].quantile(0.75)
    df['High_Aging'] = (df['Aging_Stock'] > aging_threshold).astype(int)
    inefficiency_components.append('High_Aging')

# Score totale inefficienza
if inefficiency_components:
    df['Inefficiency_Score'] = df[inefficiency_components].sum(axis=1)

    # Classifica efficienza
    def classify_efficiency(score):
        if score == 0:
                return 'Efficiente'
            elif score == 1:
                return 'Moderato'
            elif score == 2:
                return 'Inefficiente'
            else:
                return 'Critico'

        df['Efficiency_Class'] = df['Inefficiency_Score'].apply(classify_efficiency)

        print(f"Calcolato Inefficiency_Score e Efficiency_Class")
        print("Distribuzione classi efficienza:")
        for cls, count in df['Efficiency_Class'].value_counts().items():
            pct = count / len(df) * 100
            print(f"  {cls}: {count:,} ({pct:.1f}%)")

    return df


def prepare_clustering_features(df):
    """Prepara features per clustering (gestione base missing per categorie prima dell'encoding)"""
    print('\n' + '='*70)
    print('PHASE 3 - PREPARAZIONE FEATURES PER CLUSTERING')
    print('='*70)

    # Features numeriche per clustering
    numeric_features = [
        'Weight', 'Volume', 'Volume_Efficiency',
        'Storage_Days', 'Lead_Time', 'Aging_Stock',
        'Storage_Cost_Index', 'Stock_Rotation_Efficiency', 'Expiration_Risk'
    ]

    # Filtra solo features disponibili
    available_numeric = [f for f in numeric_features if f in df.columns]
    print(f"Features numeriche disponibili: {len(available_numeric)}")
    for f in available_numeric:
        print(f"  - {f} (nulls: {df[f].isna().sum()})")

    # Features categoriche per encoding
    categorical_features = ['Product_Type', 'Division', 'Warehouse',
                           'Destination_Country', 'Strategy', 'Temperature']
    available_categorical = [f for f in categorical_features if f in df.columns]

    # Encoding categoriche (Label Encoding per clustering)
    encoded_features = []
    label_encoders = {}

    for feature in available_categorical:
        # Riempio i missing con una stringa 'MISSING' prima di label-encode
        df[feature] = df[feature].fillna('MISSING')
        le = LabelEncoder()
        encoded_col = f"{feature}_encoded"
        # cast a str per sicurezza, poi a int
        df[encoded_col] = le.fit_transform(df[feature].astype(str)).astype(float)
        encoded_features.append(encoded_col)
        label_encoders[feature] = le
        print(f"Encoded {feature} -> {encoded_col} ({df[feature].nunique()} categorie, nulls now:
{df[encoded_col].isna().sum()})")

    # Combina tutte le features per clustering
    all_clustering_features = available_numeric + encoded_features

    print(f"\nTotale features per clustering: {len(all_clustering_features)}")

    return df, all_clustering_features, label_encoders


def perform_advanced_clustering(df, features):
    """Esegue clustering con K-Means e DBSCAN per identificare tipologie prodotti
      + imputazione semplice per NaN prima dello scaling"""
    print('\n' + '='*70)
    print('PHASE 4 - CLUSTERING PRODOTTI')
    print('='*70)

    if not features:
        print("Nessuna feature disponibile per clustering")
        return df, None, None, 2

    # Prepara matrice features
    X = df[features].copy()

    # Sostituisci inf con NaN per sicurezza
    X.replace([np.inf, -np.inf], np.nan, inplace=True)

    # Diagnostica missing prima dell'imputazione
    total_missing = int(X.isnull().sum().sum())
    if total_missing > 0:
        print(f"ATTENZIONE: trovati {total_missing} valori mancanti nella matrice di clustering. Procedo
con imputazione (mediana).")
        imputer = SimpleImputer(strategy='median')
        X_imputed = imputer.fit_transform(X)
        X = pd.DataFrame(X_imputed, columns=X.columns, index=X.index)
        print("Imputazione completata (strategy=median).")
    else:
        print("Nessun missing nella matrice di clustering.")

    # Standardizzazione
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    print(f"Matrice di clustering: {X_scaled.shape}")

    # 1. K-MEANS CLUSTERING
    print("\n--- K-Means Clustering ---")

    # Trova k ottimale con silhouette score
    # Garantisco che max_k sia almeno 2
    max_k = min(10, max(3, len(df)//5))
    k_range = range(2, max_k)
    silhouette_scores = []
    kmeans_models = []

    for k in k_range:
        kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
        cluster_labels = kmeans.fit_predict(X_scaled)
        try:
            score = silhouette_score(X_scaled, cluster_labels)
        except Exception as e:
            print(f"  k={k}: silhouette score non calcolabile ({e}), imposto score=-1")
            score = -1
        silhouette_scores.append(score)
        kmeans_models.append(kmeans)
        print(f"  k={k}: silhouette_score={score:.3f}")

    # Seleziona k ottimale (gestione caso tutti score -1)
    if len(silhouette_scores) == 0 or max(silhouette_scores) <= -0.5:
        optimal_k = 2
        best_kmeans = kmeans_models[0]
        best_silhouette = silhouette_scores[0] if silhouette_scores else None
        print("Impossibile trovare silhouette valida, uso k=2 di fallback.")
    else:
        best_idx = int(np.argmax(silhouette_scores))
        optimal_k = list(k_range)[best_idx]
        best_kmeans = kmeans_models[best_idx]
        best_silhouette = silhouette_scores[best_idx]

    df['KMeans_Cluster'] = best_kmeans.labels_

    print(f"K-Means ottimale: k={optimal_k}, silhouette={best_silhouette:.3f}" if best_silhouette is not None
else f"K-Means ottimale: k={optimal_k}")
    print("Distribuzione cluster K-Means:")
    for cluster_id in sorted(df['KMeans_Cluster'].unique()):
        count = (df['KMeans_Cluster'] == cluster_id).sum()
        pct = count / len(df) * 100
        print(f"  Cluster {cluster_id}: {count:,} ({pct:.1f}%)")

    # 2. DBSCAN per identificare outliers
    print("\n--- DBSCAN Clustering ---")

    # Prova diversi parametri DBSCAN
    eps_values = [0.5, 1.0, 1.5, 2.0]
    best_dbscan = None
    best_eps = None

    for eps in eps_values:
        dbscan = DBSCAN(eps=eps, min_samples=5)
        cluster_labels = dbscan.fit_predict(X_scaled)
        n_clusters = len(set(cluster_labels)) - (1 if -1 in cluster_labels else 0)
        n_noise = list(cluster_labels).count(-1)

        if n_clusters > 1 and n_noise < len(df) * 0.5:  # Evita troppo rumore
            best_dbscan = dbscan
            best_eps = eps
            print(f"  eps={eps}: {n_clusters} cluster, {n_noise} outliers")
            break

    if best_dbscan:
        df['DBSCAN_Cluster'] = best_dbscan.labels_
        print(f"DBSCAN selezionato: eps={best_eps}")

        # Conta outliers
        outliers = (df['DBSCAN_Cluster'] == -1).sum()
        print(f"Outliers identificati: {outliers} ({outliers/len(df)*100:.1f}%)")
    else:
        df['DBSCAN_Cluster'] = 0  # Fallback
        print("DBSCAN non ha prodotto risultati validi, uso cluster unico")

    # 3. ANALISI DIMENSIONALE con PCA
    print("\n--- Analisi Componenti Principali ---")
    pca = PCA(n_components=min(3, len(features)))
    X_pca = pca.fit_transform(X_scaled)

    print("Varianza spiegata per componente:")
    for i, var_ratio in enumerate(pca.explained_variance_ratio_):
        print(f"  PC{i+1}: {var_ratio:.3f} ({var_ratio*100:.1f}%)")

    # Aggiungi componenti principali al dataframe
    for i in range(pca.n_components_):
        df[f'PC{i+1}'] = X_pca[:, i]

    return df, X_scaled, best_kmeans, optimal_k


def analyze_product_clusters(df, optimal_k):
    """Analizza profili dei cluster per identificare tipologie prodotti e inefficienze"""
    print('\n' + '='*70)
    print('PHASE 5 - ANALISI PROFILI CLUSTER')
    print('='*70)

    cluster_profiles = []

    for cluster_id in range(optimal_k):
        cluster_data = df[df['KMeans_Cluster'] == cluster_id]

        if len(cluster_data) == 0:
            continue

        # Statistiche numeriche
        profile = {
            'Cluster_ID': cluster_id,
            'Size': len(cluster_data),
            'Percentage': len(cluster_data) / len(df) * 100
        }

        # KPI medi
        numeric_kpis = ['Weight', 'Volume', 'Storage_Days', 'Lead_Time',
                       'Aging_Stock', 'Inefficiency_Score']

        for kpi in numeric_kpis:
            if kpi in cluster_data.columns:
                profile[f'Avg_{kpi}'] = cluster_data[kpi].mean()

        # Calcola KPI derivati
        if 'Storage_Cost_Index' in cluster_data.columns:
            profile['Avg_Storage_Cost'] = cluster_data['Storage_Cost_Index'].mean()

        if 'Expiration_Risk' in cluster_data.columns:
            profile['Avg_Expiration_Risk'] = cluster_data['Expiration_Risk'].mean()

        # Distribuzione classi efficienza
        if 'Efficiency_Class' in cluster_data.columns:
            profile['Efficiency_Distribution'] = cluster_data['Efficiency_Class'].value_counts().to_dict()

        # Top prodotti nel cluster
        if 'Product_Type' in cluster_data.columns:
            profile['Top_Products'] = cluster_data['Product_Type'].value_counts().head(3).to_dict()

        # Top destinazioni
        if 'Destination_Country' in cluster_data.columns:
            profile['Top_Destinations'] = cluster_data['Destination_Country'].value_counts().head(3).to_dict()

        # Strategia predominante
        if 'Strategy' in cluster_data.columns:
            profile['Strategy_Mix'] = cluster_data['Strategy'].value_counts().to_dict()

        cluster_profiles.append(profile)

        # Stampa profilo cluster
        print(f"\nCLUSTER {cluster_id} ({profile['Size']} prodotti, {profile['Percentage']:.1f}%):")

        if 'Avg_Inefficiency_Score' in profile:
            score = profile['Avg_Inefficiency_Score']
            status = "CRITICO" if score >= 2.5 else "INEFFICIENTE" if score >= 1.5 else "MODERATO" if score
>= 0.5 else "EFFICIENTE"
            print(f"  Status: {status} (score: {score:.2f})")

        if profile.get('Top_Products'):
            main_product = list(profile['Top_Products'].keys())[0]
            print(f"  Prodotto principale: {main_product}")

        for kpi in ['Avg_Storage_Days', 'Avg_Lead_Time', 'Avg_Aging_Stock']:
            if kpi in profile:
                print(f"  {kpi.replace('Avg_', '')}: {profile[kpi]:.1f}")

    return cluster_profiles


def identify_supply_chain_inefficiencies(df, cluster_profiles):
    """Identifica cluster critici e pattern di inefficienza"""
    print('\n' + '='*70)
    print('PHASE 6 - IDENTIFICAZIONE INEFFICIENZE')
    print('='*70)

    # Ranking cluster per inefficienza
    ranked_clusters = sorted(cluster_profiles,
                    key=lambda x: x.get('Avg_Inefficiency_Score', 0),
                    reverse=True)

    # Identifica cluster critici e efficienti
    avg_inefficiency = np.mean([p.get('Avg_Inefficiency_Score', 0) for p in cluster_profiles])

    critical_clusters = [p['Cluster_ID'] for p in ranked_clusters
                  if p.get('Avg_Inefficiency_Score', 0) >= avg_inefficiency * 1.3]

    efficient_clusters = [p['Cluster_ID'] for p in ranked_clusters
                  if p.get('Avg_Inefficiency_Score', 0) <= avg_inefficiency * 0.7]

    print(f"Score inefficienza medio: {avg_inefficiency:.2f}")
    print(f"Cluster critici: {critical_clusters}")
    print(f"Cluster efficienti: {efficient_clusters}")

    # Analisi pattern specifici
    print("\n--- Pattern di Inefficienza Identificati ---")

    patterns = []

    # Pattern 1: Alto costo storage
    high_storage_clusters = [p['Cluster_ID'] for p in cluster_profiles
                    if p.get('Avg_Storage_Days', 0) >
                      df['Storage_Days'].quantile(0.8) if 'Storage_Days' in df.columns]
    if high_storage_clusters:
        patterns.append(f"Cluster ad alto storage: {high_storage_clusters}")

    # Pattern 2: Lead time elevati
    high_leadtime_clusters = [p['Cluster_ID'] for p in cluster_profiles
                    if p.get('Avg_Lead_Time', 0) >
                      df['Lead_Time'].quantile(0.8) if 'Lead_Time' in df.columns]
    if high_leadtime_clusters:
        patterns.append(f"Cluster con lead time elevati: {high_leadtime_clusters}")

    # Pattern 3: Rischio scadenza
    high_expiration_clusters = [p['Cluster_ID'] for p in cluster_profiles
                      if p.get('Avg_Expiration_Risk', 0) > 2.0]
    if high_expiration_clusters:
        patterns.append(f"Cluster a rischio scadenza: {high_expiration_clusters}")

    for pattern in patterns:
        print(f"  {pattern}")

    return critical_clusters, efficient_clusters, patterns
```