

Politecnico di Torino

Department of Management and Production Engineering

Master's degree in Management Engineering

Improving demand and inventory forecast with data analytics techniques in a real manufacturing business case



Supervisor

Tania Cerquitelli

Candidate

Matteo Accarrino

Academic Year 2018-2019

*Ai miei genitori,
a Lino,
a chi mi guarda da lassù.*

*”É impossibile vivere senza fallire in qualcosa, a meno che non
vivate in modo così prudente da non vivere del tutto.
In quel caso, avrete fallito in partenza.”*

[J.K.Rowling]

Improving demand and inventory forecast with data analytics techniques in a real manufacturing business case

Matteo Accarrino

Supervisor:

Tania Cerquitelli

Abstract

The manufacturing industry has always had to deal with a contraposition in the context of the inventory management. On the one hand, it is necessary ensure the financial efficiency by trying to build up smaller inventories, in order to reduce costs of their management and to avoid company's assets slow and ineffective turn over. On the other hand, it is essential make sure that there is an operational continuity and, as a consequence, an high level of service.

Therefore, companies are called to define a right and winning balance. In this fundamental task, data analytics techniques can help them, becoming a key factor in decision-making process and, more generally, an indispensable strategic weapon. Particularly, they may be a value added support to demand sizing and subsequent production forecasting, making them more effective and robust, translating, hence, into a competitive advantage.

The purpose of this study is precisely to define, in a real manufacturing business case, the best data-based approach to demand and inventory forecasting. It starts from data exploration aimed at identifying the key insights that can lead to a better forecasting approach, taking into account also cost impact and periodicity of employment. Then a product clustering is carried out, to define clusters with similar behaviors, that must be treated

separately and independently. Finally, the forecast modelling is implemented via 2 alternative and parallel methodologies (Multilinear Regression and Random Forest Tree regression) and results are compared to identify the best combination, leading to the highest accuracy for each cluster.

Contents

1	Introduction	11
1.1	Problem statement	11
1.2	Objectives	12
1.3	Analytical tools used in this study	12
1.3.1	Knime Analytics Platform	12
1.3.2	Microsoft Power BI	13
1.4	Contents overview	14
2	Literature review	16
2.1	Inventory Management	16
2.1.1	The relationship with the firm performance	21
2.2	Inventory Management analytics	27
3	Dataset review	33
3.1	Data gathering	33
3.1.1	Stock	34
3.1.2	Delivered Products	35
3.1.3	Gross requirements	36
3.1.4	Items descriptions	37
3.2	Main measures statistics	39
3.2.1	Gross Requirements Statistics	39
3.2.2	Stock Quantity Statistics	45
3.2.3	Delivered Quantity Statistics	47
4	Data Preparation	52
4.1	Explorative analysis	52

4.1.1	Cost based ABC analysis	53
4.1.2	Seasonality and time usage	59
5	Modeling	63
5.1	Cluster analysis	63
5.1.1	K-Means clustering	64
5.1.2	K-Means implementation	64
5.1.3	Clustering findings	70
5.2	Forecasting	73
5.2.1	Approach to forecast	74
5.2.2	Multilinear Regression	77
5.2.3	Random Forest Regression	78
5.2.4	Model evaluations tools	80
5.2.5	Cluster 0	81
5.2.6	Cluster 1	84
5.2.7	Cluster 2	86
5.2.8	Cluster 3	89
6	Conclusions	92

List of Figures

1.1	Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms (as of Nov 2018)	13
1.2	Gartner 2019 Magic Quadrant for Analytics and Business Intelligence Platforms (as of Jan 2019)	14
2.1	Inventories and the flow of items	17
2.2	ABC curve: percentage of value versus percentage of items.	20
2.3	Sales trend during the firm life cycle [19]	23
2.4	These are two examples of the inventory cycle, with changes in the level of stocks behind changes of income. The two charts differ in the amplitude of the cycles as the sequence develops.[20]	24
2.5	The ELI denotes the studentized deviation of a firm's inventory holdings from its peers within the same industry.	25
2.6	Supply chain benefits achieved using big data analytics[1].	30
2.7	Problems about the use of big data analytics[1]. .	31
2.8	How tools and technology support the use of big data analytics[1].	32
3.1	Workflow to read multiple Excel files and append them in a single file.	34
3.2	Example of the actual gross requirement quantity.	38
3.3	Data extraction, reading and integration pipelines	39
3.4	Total Requirements quantity trend	40

3.5	Total Requirements Statistics	40
3.6	Requirement Quantities Quartiles	41
3.7	Requirements Statistics of the 1st quartile - high .	41
3.8	Requirements Trend of the products in the 1st quartile - high	42
3.9	Requirements Statistics of the 2nd quartile - medium high	42
3.10	Requirements Trend of the products in the 2nd quartile - medium high	42
3.11	Requirements Statistics of the 3rd quartile - medium low	43
3.12	Requirements Trend of the products in the 3rd quartile - medium low	43
3.13	Requirements Statistics of the 4th quartile - low .	43
3.14	Requirements Trend of the products in the 4th quartile - low	44
3.15	Total Stock quantity trend as per initial data extraction	45
3.16	Total Stock quantity trend after data cleaning . .	46
3.17	Total Stock quantity Statistics	47
3.18	Total Delivered quantity trend	47
3.19	Total Delivered quantity Statistics	48
3.20	Delivered quantities Quartiles	48
3.21	Delivered Products Statistics of the 1st quartile - high	49
3.22	Delivered Trend of the products in the 1st quartile - high	49
3.23	Delivered Products Statistics of the 2nd quartile - medium high	49
3.24	Delivered Trend of the products in the 2nd quartile - medium high	50
3.25	Delivered Products Statistics of the 3rd quartile - medium low	50

3.26	Delivered Trend of the products in the 3rd quartile - medium low	50
3.27	Delivered Products Statistics of the 4th quartile - low	51
3.28	Obsolete Products distribution across Delivered Quar- tiles	51
4.1	Workflow implemented to conduct the cost based ABC analysis.	54
4.2	Portion of graph that shows the number of items' percentage of their grand total.	56
4.3	Portion of graph that shows the items' cost per- centage of their grand total.	56
4.4	Monthly gross requirements trend.	60
4.5	Daily gross requirements weight over the week trend in the working weeks	61
4.6	Weekly and yearly usage relationship.	61
5.1	Loop to compute Elbow Method.	67
5.2	Calculation of sum of squared errors (SSE) for each iteration.	68
5.3	Finding of the appropriate number of clusters. . .	69
5.4	Scatter plot of the SSE for all clusterings.	70
5.5	Scatter plot matrix with cluster evidence - key di- mensions relationship	72
5.6	Cluster Cardinality and MRP classification com- parison	73
5.7	Obsolete Products presence in each cluster	73
5.8	Building of items' past N values vector for gross requirements.	74
5.9	Recursive loop carried out to predict multiple days after the next one.	76
5.10	Building of items' past N values vector for stock values.	77
5.11	Random Forest algorithm schema.	80

5.12	Example of results of predictions for Cluster 0 items both for stock and gross requirements.	83
5.13	Example of results of predictions for Cluster 1 items both for stock and gross requirements.	85
5.14	Example of results of predictions for Cluster 2 items both for stock and gross requirements.	88
5.15	Example of results of predictions for Cluster 3 items both for stock and gross requirements.	91

List of Tables

2.1	Industry of companies involved in the survey. . . .	30
3.1	Stock data dictionary.	34
3.2	Delivered data dictionary.	35
3.3	Delivered processing method.	36
3.4	Gross requirements data dictionary.	37
3.5	Part description data dictionary.	37
4.1	Stock - ABC classification's results over 3, 6 and 12 last months.	57
4.2	Stock-Requirements ABC Matrix.	58
5.1	Correlation Matrix.	66
5.2	SSE for each iteration.	68
5.3	SSE delta sorting.	69
5.4	Clusters' centers description.	70
5.5	Portion of an items' past 7 values vector.	75
5.6	Accuracy performances of cluster 0 predictions. .	82
5.7	Statistics of cluster 0 predictions.	82
5.8	Accuracy performances of cluster 1 predictions. .	84
5.9	Results with and without model mismatch for stock values predictions.	86
5.10	Statistics of cluster 1 predictions.	86
5.11	Accuracy performances of cluster 2 predictions. .	86
5.12	Results with and without model mismatch for stock values predictions.	87
5.13	Statistics of cluster 2 predictions.	87
5.14	Accuracy performances of cluster 3 predictions. .	89

5.15	Accuracy performances of cluster 3 predictions. .	89
5.16	Statistics of cluster 3 predictions.	90

Chapter 1

Introduction

1.1 Problem statement

The manufacturing industry is historically calling for better forecasting all along the supply chain blocks, in order to drive higher quality of service and financial efficiency. Better demand, supply and inventory control would lead to better overall company performance and customers and stakeholders fidelization, too. Getting more robust forecast could definitely turn into a competitive advantage for a company who is highly exposed to supply chain challenges.

When looking at inventory management, finding a sweet spot in constantly balancing cost and service implications is key to company success. Production forecasting, essential to company flourishing, is a continuous exercise, constantly influenced by market trends, customer behaviour and company strategy. Finally a vertically integrated supply chain, where suppliers governance guarantees the highest operational standards, can become a competitive advantage and is needed to excel.

The study has been carried out in collaboration with a global manufacturer, around its core supply chain processes. This company provided a large collection of supply chain data available through the ERP and Business Intelligence infrastructure, along with the domain expertise required to interpret the data and the

final results, and the provision of feedback and support as needed.

1.2 Objectives

The purpose of this study is to define, in a real complex manufacturing business case, the best data-based approach to demand and inventory forecasting, leveraging state of the art data mining techniques. Starting from data made available by the company, the study will focus on identifying the key insights that can lead to a better forecasting modeling, taking into account products peculiarities that may arise from data analysis, so that the overall forecast accuracy is maximized.

1.3 Analytical tools used in this study

In order to carry out the study a mix of tools has been used, so that scalability, performance and reproducibility of the study could be guaranteed.

1.3.1 Knime Analytics Platform

Most of this study has been carried out using Knime. Indicated as one of the four leading analytical platforms for data science and machine learning by Gartner in 2019 (Figure 1.1), Knime is a free and open-source data analytics, reporting and integration platform. In order to enable users to conduct their own analysis, it provides various machine learning and data mining components, including preprocessing (ETL: Extraction, Transformation, Loading), modeling, data analysis and visualization without, or with only minimal, programming. Moreover, Knime makes use of extensions to add plugins and additional functionalities. Each of its components is represented by a node. Users visually create workflows assembling them. In this way, they can selectively execute some or all analysis steps, inspect the results, models, and



Figure 1.1: Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms (as of Nov 2018)

interactive views. Furthermore Knime core-architecture ensures processing of large data volumes. User-friendliness, attractive graphical interface, expandability and modularity are some of the features that have guaranteed Knime’s success and fast growing, and that led me to choose it as core platform in my analysis.

1.3.2 Microsoft Power BI

For 12 consecutive years, Gartner has recognized Microsoft as a leader in analytics and business intelligence and the company still arises as one of the top players in the BI space in 2019 (Figure 1.2). Microsoft Power BI is a business analytics service aimed at delivering insights that enable faster, informed decisions. It allows data transformation into visuals that help communicating more effectively the key insights as well fast data exploration through a visual interactive experience. Its proven scalability, leveraging also collaboration features and built-in security governance, made it an industry reknown platform widely adopted across mid and

large size companies.



Figure 1.2: Gartner 2019 Magic Quadrant for Analytics and Business Intelligence Platforms (as of Jan 2019)

1.4 Contents overview

This study covers several topics related to the subject matter, in order to build a comprehensive corpus that can introduce and sustain its thesis and conclusions.

The dissertation starts with a literature review, covering the key concepts related to inventory management and its links to firm performances.

A chapter on dataset review follows, presenting the data sources, the data types and their definition per each available measure. The data are also presented from a statistical standpoint, highlighting key dynamics and characteristics that will be further explored in the course of the study.

Exploring data is the next step, so that some business relevant insights can be already driven and few hypothesis, such as need for

product clustering before stepping into regression and forecasting, are set out and validated. This is done conducting a classical ABC analysis on stock value and crossing results with an ABC analysis on requirements value. Furthermore the periodicity is analyzed and turned into a variable for following usage.

The most important phase is then modeling, which follows a 2 step approach: unsupervised learning is leveraged to identify product clusters which share similar behavior, so that supervised learning through regression can be applied separately and independently, reaching a higher accuracy. Regression is then applied via 2 parallel methodologies (Multilinear and Random Forest Tree) and results are reported to identify the best combination leading to the highest accuracy for each cluster.

Last chapter is devoted to presenting the study results and driving the conclusions.

Chapter 2

Literature review

2.1 Inventory Management

In 1991 the Council of Logistics Management, a trade organization based in the United States, defined the inventory management as: *"the process of planning, implementing, and controlling the efficient, effective flow and storage of goods, services, and related information from point of origin to point of consumption for the purpose of conforming to customer requirements"*.

The Inventory Management, therefore, is a process responsible for planning and controlling inventory from the raw material stage to the customer. It results from production and supports it, for this reason the two can't be managed separately and must be coordinated[6].

The inventory is build up in raw materials, work-in-process, and finished goods considering the flow of elements into, through, and out of a firm (Figure 2.1):

- raw materials are basic materials that have not entered the production process. They are used to produce goods, finished products, or intermediate materials which are feedstock for future finished products;
- work-in-process (WIP) is the label for items that have entered the manufacturing process and are being worked on or waiting to be processed in a queue or a buffer storage;

- finished goods are goods that have completed the manufacturing process but have not yet been sold or distributed to the end user.

This classification also provides for elements used in production that do not become part of the product (maintenance, repair, and operational supplies).

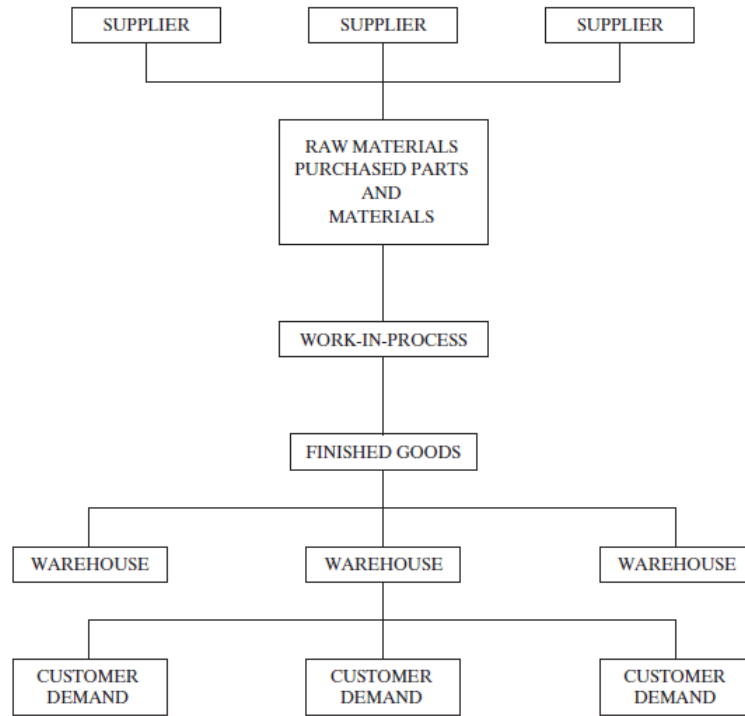


Figure 2.1: Inventories and the flow of items

All business require inventories and the importance of their managing can't be underestimated[2].

Inventories have functions in regard to productive and distributive operations in firms[7]. On the one hand they ensure to meet customer's demand and to provide along the logistics chain materials and supplies as inputs to the production process; on the other hand they allow to decouple production phases sequentially linked and operating at different speeds and they guarantee availability to reduce the time of the order-delivery cycle. In

this way there is continuity and flexibility in the production plan. Considering this goal, the build up of an inventory also allow to cope with uncertainty (e.g. changes in delivery times of incoming components or demand) and secure acceptable service levels when balancing supply and demand[21]. In fact, if supply met demand exactly, there wouldn't be need for building up an inventory: this happens when the demand is predictable, static and relatively constant over a long time period. If this is so, goods can produced on a line-flow basis and they are delivered to the customer at the rate the customer needs them. However, firms face uncertainty due to a demand that isn't constant and deterministic. In this context, stockout or overstock situations may be occur: if demand or lead time is greater than forecast, a stockout will occur, while if demand or lead time is less than forecast, an overstock will occur. They are cases that a firm can usually easily avoid if there is certainty, but in real cases they must be taken into account.

Thus, it is necessary to build up an inventory in order to bring more flexibility and responsiveness in a firm[7] against the changes or fluctuations in demand and production[6] and its management must be effective. For this reason, inventories are classified according to the function they perform[6]:

- anticipation inventories are built up to support a certain level of production and to reduce the costs of changing production rates;
- fluctuation inventories cover unpredictable fluctuations in supply and demand or lead time and prevent disruptions in manufacturing or deliveries to customers. They represent the safety stock;
- lot-size inventories deplete gradually as customers' orders come in and is replenished cyclically when suppliers' orders are received. Their aim is to reduce shipping, clerical, and setup costs;

- transportation inventories are movement inventories: they exist because of the time needed to move items from one place (e.g. plant, distribution center) to another (e.g. customer).

Moreover, inventories obviously affect economics in firms. They are a part of a firm's total assets (they usually represent from 20% to 60% of total value[6]), so they represent a significant cost factor. Supply, manufacturing, warehousing, and transportation investments and activities indeed are determiners of operating costs increase and of profits decrease: these costs increase with the growth of the inventory.

Among these costs there are risk costs that are due to the risks in carrying inventory, like damage, pilferage, deterioration or obsolescence. This last situation occurs when an item is not used for a period of time. This leads to accumulation of the dead parts that can't be sold back to the suppliers as it is too late as per contract agreement[5]. The inventory management should be as efficient as possible.

In addition to direct costs, it is also important to not forget indirect costs: for instance, if a stockout occurs there is a direct impact, the income foregone, but also indirect effects like the firm's loss of image and of confidence. For this reason, the inventory management must maximize customer service, that is the ability of a company to satisfy the needs of customers at the right time and with the right products/services[29].

However, as asset, inventory's effective exploitation and the revenues, that come from the sale of goods or services, ensure the improvement of the cash flow and of the return on investment.

Given the considerable high costs of carrying inventory, besides managing inventory at the aggregate level, it must also be managed at items level. Indeed, to have better control at a reasonable cost, it is helpful to classify the items according to their importance, measured usually through annual euro/dollar usage (other measures can be either unit cost or scarcity of material

too)[6]. This control is carried out by the ABC inventory. It is a classification system allowing different level of control based on the items' importance. The ABC inventory is based on the deduction, resulting from Pareto's law, that the relationship between the percentage of items and the percentage of annual dollar usage follows a pattern which allows to define three groups(Figure 2.2):

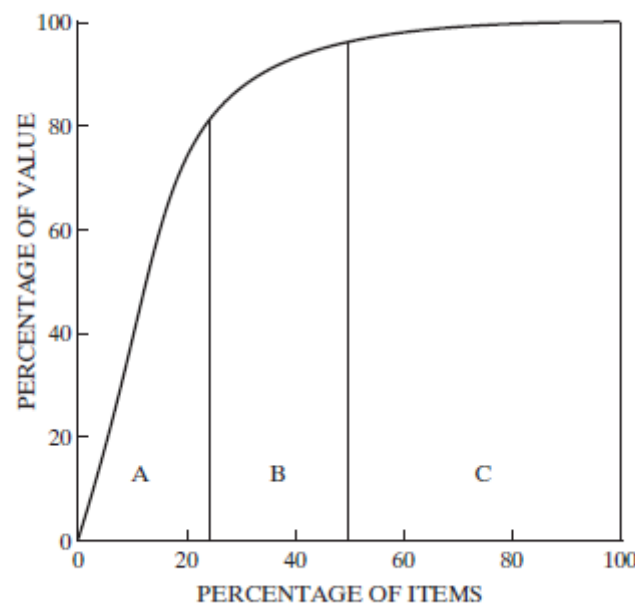


Figure 2.2: ABC curve: percentage of value versus percentage of items.

- group A - about 20% of items resulting in about 80% of the dollar/euro usage;
- group B - about 30% of items resulting in about 15% of the dollar/euro usage;
- group C - about 50% of items resulting in about 5% of the dollar/euro usage;

The percentages should not be taken as absolute because they are approximate. According to the group to which each items

belong, there is a different degree of management and control. A items have high priority and a tight control that include regular and frequent review, accurate records and demand forecasts with steady review and close follow-up and expediting to reduce lead time B items have medium priority involving normal controls, regular attention, and normal processing. Instead, C items have lowest priority with simplest controls and records, and a periodic review system. This would indicate that it is advisable to have extra stock of C items which adds little to the total inventory value but they are fundamental in order to avoid shortage.

In conclusion, therefore, there are both benefits and costs to having inventory. Thus inventory management optimization depends on two drivers: total cost minimization (efficiency) and net revenue maximization (effectiveness). These optimizations are performed subject to constraints on meeting demand and delivery response times, and satisfy customer service[25].

What follows provides a detailed analysis of the impact on the firm performance.

2.1.1 The relationship with the firm performance

Inventory management has a significant influence on supply chain and firm performance[25], as already explained. Due to its influence as driver of performance and the costs that could be incurred if it is not managed optimally, firms undertake numerous initiatives to improve inventory management efficiency and effectiveness[15]. In this category there is the exploitation of information technology with supply chain softwares. The importance of these techniques is self evident and widely accepted. The majority of success stories in operations management indeed show that the firms' employment of the previous techniques led to increased market share, higher profitability and greater product quality.[4]

In order to ensure an improvement it should be specified that reducing inventories could be not a good practice. Managers can be brought to apply this policy because inventory holding costs

money. Obermaier and Donhauser's studies[21] demonstrate that firms with the lowest inventory have the worst performance (and vice versa), interpreting performance as a function of inventory, while the low-performing firms carry the least inventory, whereas high performing firms have the highest stocks, interpreting inventory as a function of performance. This results from analysis conducted about the inventory performance between 1989 and 2004 of a sample of firms assigned to the Standard Industrial Classification (SIC) manufacturing division in Germany, as a major European economy. To investigate their hypotheses they run several time series regressions. They found a relationship between inventory and firm performance: increasing inventories lead to increasing financial performance (and vice versa). Moreover, results reflect a certain inventory behaviour in times of low financial performance: if a firm have to face financial trouble, it may be forced to reduce inventories in order to ensure an increment of short-term liquidity, but this will not help the firms in better times, when it is necessary to have inventories to production and serve their customers. Many firms indeed go bankrupt. Low inventories also make it much more difficult to manage business processes cost-efficiently and, as results suggest, a reduction in inventory turns out to be a wrong choice considering the negative effects on business performance.

However, a fair level of inventory can't be prescribed regardless. The relationship between inventory and firm performance can vary with organizational life cycle stage (Figure 2.3), because the decision of inventory depends, like other decisions, not only on its forecasted costs and benefits but also on the environment that the organization is confronting.

This is what Elsayed and Wahba deduced through econometric analysis[14]. They referred to a sample of 84 Egyptian firms coverage eighteen industrial sectors with total number of observations of 504, from 2005 to 2010. The study took into account the organization performance, measured by the return on asset

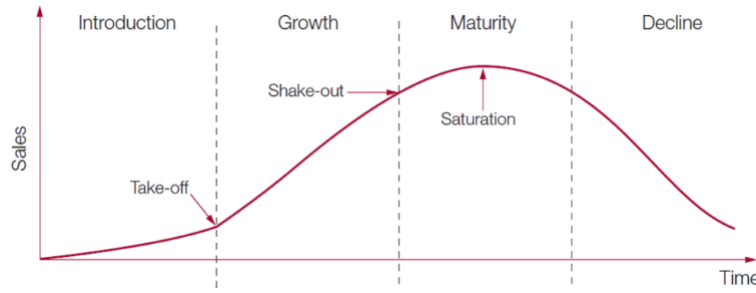


Figure 2.3: Sales trend during the firm life cycle [19]

(ROA) because this indicator reflects operating results. It was assessed in relation to the inventory to sale ratio and other associated variables as the organization size or the financial leverage. The results suggest that the relationship between inventory and organization performance is negative in the initial growth stage and the maturity stage: in the initial growth stage indeed firms have fluctuated demand and relatively small ordering quantities, while in the maturity stage demand and growth rate become more stable and firms put more emphasis on cost control to improve efficiency. Therefore holding more inventories affects performance negatively. Instead the relationship is positive in the rapid growth stage and the revival stage: in the rapid growth firms need more inventory to stimulate demand and to increase sales growth by increase service levels of existing products or by the introduction of new products; in the revival stage there is the re-inventing business, so firms need to change their product-market strategy with other products in order to stimulate a new growth and rebuild their market share and profitably. Therefore it is expected that inventory in these case affects performance positively.

What Elsayed and Wahba found was supported by economists which long recognized the important role that inventories play in business cycles[20](Figure 2.4).

The relationship between firm performance and inventory management has involved many other studies, too.

Capkun[4]confirmed that improving a firm's inventory performance guarantees better financial performance measured both at

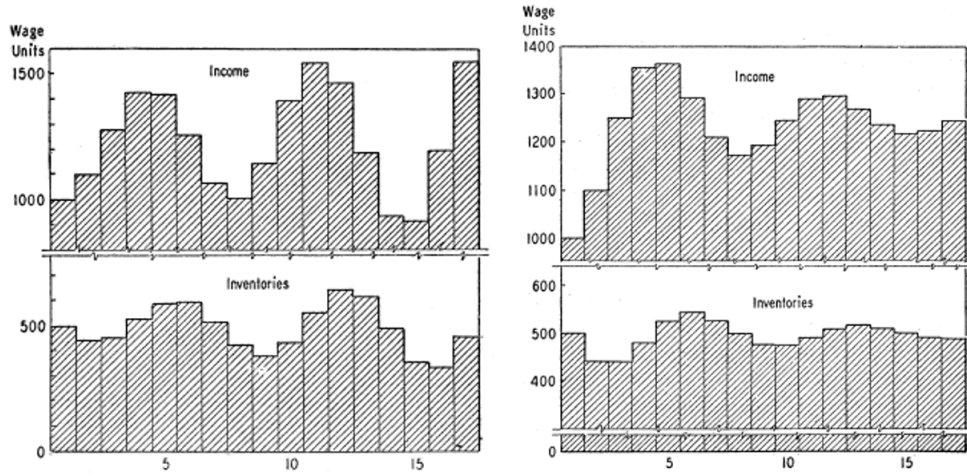


Figure 2.4: These are two examples of the inventory cycle, with changes in the level of stocks behind changes of income. The two charts differ in the amplitude of the cycles as the sequence develops.[20]

gross (gross profit is the profit makes after deducting the costs associated with making and selling its products, or the costs associated with providing its services) and operating levels (operating profit is the profit earned from a company's ongoing core business operations, excluding deductions of interest, taxes and any profit earned from the firm's investments). He used as firm's inventory performance indicator the inventory to sale ratio that measures the amount of inventory compared to the number of sales fulfilled. It is a strong indicator of prevailing economic conditions: a lower value is better. To test this positive correlation, Capkun analyzed a sample of US manufacturing firms from 1980 to 2005, excluding observations without data available on raw materials inventory, WIP inventory, or finished goods inventory. The choice to exclude these observations is due to the identification also of the potentially differential performance effects of raw materials, work-in-process, and finished goods inventories. In fact, his results show that there is a strong correlation between inventory performance and financial performance, but the strength of the correlation differs between inventory types. There are difference between these

three types of inventory: for instance, raw materials are lower in unit value as compared with finished goods and demand for the latter is more uncertain than that for the former. Thus, it is conceivable that such differences may have different effects on firm performance, so that the effect of total inventories on firm performance can be decomposed into distinct performance effects of the three types of inventory.[15]. Regressions test conducted by Capkun proved that raw materials inventory performance has the highest correlation with all financial performance measures, while WIP inventory is more highly correlated with the gross financial performance and finished good inventory performance has a stronger correlation with operating financial performance.

The conclusions he came to are in line with those reached by Eroglu[15]. His study examined data of 885 firms from 27 US manufacturing industries, from 2003 to 2008 with the use regressive models: ROS, the return on sales for firm, was the dependent variable and measured firm performance; instead inventory management performance was measured using the ELI, an indicator introduced by Eroglu. The ELI measures a firm's inventory leanness from the size-adjusted within-industry average inventory level(Figure 2.5).

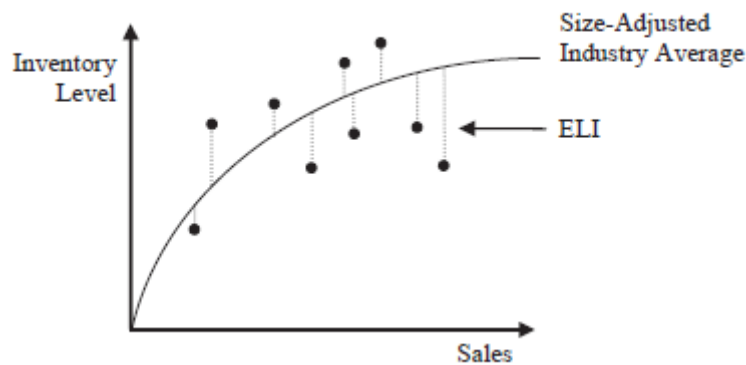


Figure 2.5: The ELI denotes the studentized deviation of a firm's inventory holdings from its peers within the same industry.

Eroglu's findings confirm that the performance effect of raw

material inventory has the greatest effect of financial performance among all inventory types. The given explanation is the intertemporal interactions between raw materials inventory and other inventory types. For example, the availability of raw materials defines the feasibility of production schedules and replenishment schedules. As a consequence WIP inventory and finished goods inventory levels can be seen as a function of raw materials inventory. A shortage in this last inventory can cause a shortage in finished goods inventory. Hence, raw materials inventory may affect firm performance both directly and indirectly through finished goods levels.

In the context of the relationship which have been discussed up to now, the cost of capital and its moderating rule should be considered. Firms need to consider this cost before making any decision on financing or investing in any assets. If a firm want to improve its value and achieve a viable financial soundness, there is a need for the firm to raise capital, but the cost of capital must be lower than the cash flows generated through firm's operations. Therefore, it is important to identify cost of capital as a variable that influences firm performance[2].

However, the statistical analysis over the years have showed that inventory performance may be a decisive strategic factor both in the short- and in the long-term. At the same time, a firm's strategic choice can simultaneously impact inventory and financial performance: for example, firms strategically positioning themselves to maximize their customers service level must hold greater inventories (both in quantity and variety), resulting in a positive correlation between inventory levels and financial performance[4].

After what has been said, therefore the operations management literature prescribes a managerial focus on operations performance and in particular on in inventory performance in order to create significant value for firms.

2.2 Inventory Management analytics

As discussed so far, the inventory management is a double-edged sword[28]: on one hand the economic and financial aspect must be considered with the cost associated with the build up of too big inventories, while on the other hand the operational side mustn't been forgotten for avoiding stockouts and production block injuring the service level. As core of the supply chain, inventory management deserves more attention.

However, it is quite complicated to keep the inventory management good absolutely due the uncertainty and all related risks which afflict supply chain.

This situation leads manufacturing companies to pay attention to data that inund them, encouraging new ways to produce, organize and analyze data[17]. The aim is to capitalize on their exploitation as a means for gaining a competitive advantage[8], in order to win with them[16]. In fact, the inventory management is a fertile area for the application of analytics techniques[26], whose power can fine-tune it.

Data analytics with “big data” are analytic functions which which could be adopted by companies resulting in an indispensable strategic weapon [11]. There are a lot of ways the supply chain and, specifically, the inventory management, can benefit from using predictive analytics[13]. Some of these include:

- identifying real-time patterns and behaviors, providing manufacturers as a whole insight;
- improving of the forecasts;
- identifying hidden inefficiencies to capture greater cost savings;
- tracking and analyzing data accurately;
- improving overall operations, ensuring manufacturers to take faster and better actions and making every area of operations

more efficient;

- expense optimization, with lower inventory and operations costs and a quicker response times;
- preventing defects, disruptions and others issues, reducing downtime.

Therefore, generally, they are essentials tools to improve overall operations and reducer risk and costs, helping companies to take better decisions in relation to material flow in the supply chain become tangibly better—financially and/or operationally[11].

However, the degree to which data can be used is largely determined by their quality[22]. Poor quality data affect business decisions, promoting tangible and intangible losses[17]. In fact, studies [3][23] have demonstrated that poor quality data is costly. They have estimated the costs of poor data quality to be as high as 8% to 12% of revenues and may generate up to 40 % to 60 % of expenses. Thus, the critical rule, played by data in companies decision-making process[11], has to press manufacturers an increased awareness and sensitivity to the needs for high quality data products[17].

Among analytics techniques, which can be exploited, it is possible to distinguish descriptive and predictive analytics models[26]. Descriptive analytics consist of data-visualization tools to provide real-time information regarding every single items and process in the supply chain. Instead, predictive analytics, on which the study will focus, allow forecasting on three levels, strategic, tactical, and operational, each of which, in its own way, impacts on the planning process in supply chains, influencing important decisions such as those relating to capacity planning, production planning, and, generally, to the inventory management. Predictive analytics in supply chains makes predictions based on past data and provides answers on what will be happening and, consequently, enables to anticipate the future rather than having to react only after the problems have occurred[11].

Several methods are employed to forecast. In general, they belong to two categories: the econometrics models and the machine learning methods.

Usually, forecast steps are preceded by finding inventory classification, through involve clustering techniques. The aim is to group items based on characteristic features[9], so that there is homogeneity within clusters and heterogeneity between clusters. In this way, classification system allows companies to manage differently the various clusters, conducting separate analyses and predictions. A model of classification is typically provided by the traditional ABC method, which has been treated in the previous section, but clusters improve results deriving from its implementation. ABC method considers only one classification criterion (cost usage), while clustering takes into account even more criteria , such as lead time, availability, certainty of supply, criticalness, periodicity of usage etc, so efficient groups with a greater and better characterization are carried out.

However, with regards to the results achieved in real cases, surveys conducted by *Accenture*[1], concerning 1014 companies (Table 2.1), reveal that, effectively, the exploitation of analytics has lived up to its promise with the the analyzed companies which have relied on them. They increase supply chain efficiency, improve customer service and demand fulfillment, assure faster and more effective reaction time to issues, and promote integration across the supply chain. The benefits deriving from the exploitation of the data analytics are shown in Figure 2.6.

Actually, the survey has revealed that data analytics in the supply chain is not widespread or well coordinated across companies. Although benefits big data can assure are well known, surveyed companies have faced difficulties in adopting data analytics techniques and have yet to understand how to use them in the best way to improve their performances. The main obstacles to adoption and to prevent companies from realizing the benefits of big data analytics (Figure 2.7) have been identified

Industry	Count	Percent
Electronics & High Tech	130	13%
Consumer Good & Services	129	13%
Industrial Equipment	126	12%
Banking	125	12%
Retail	123	12%
Communications	104	10%
Health Providers	82	8%
Energy	76	7%
Chemicals	65	6%
Utilities	51	5%
Others	3	0%

Table 2.1: Industry of companies involved in the survey.



Figure 2.6: Supply chain benefits achieved using big data analytics[1].

mainly in the large investment required for deploying and using analytics and in the security issues. Other reasons are privacy issues and companies shortcomings such as absence of application

cases, limited support, and lack of in-house capability.

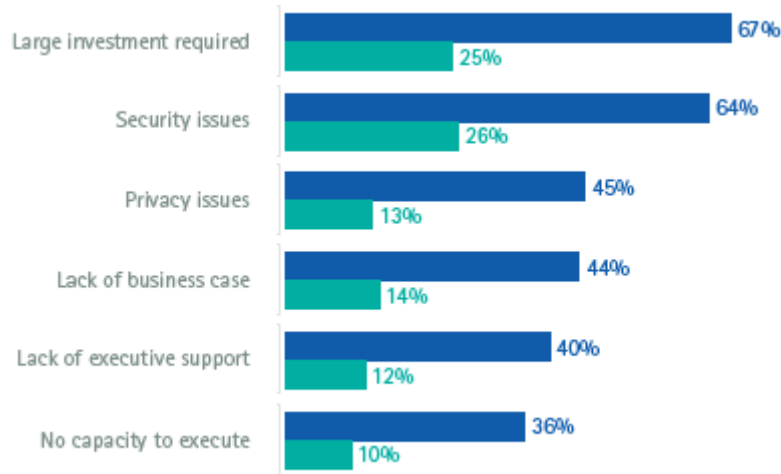


Figure 2.7: Problems about the use of big data analytics[1].

Moreover, the survey reveals that only 43% of companies rely on an company-wide data analytics capability that includes sophisticated tools to detect, process, and produce key information for the supply chain(Figure 2.8). Indeed, it would be more appropriate to connect supply chain forecasting and modeling tools to all other business aspects because the value to which company aspires isn't ensured focusing only on supply chain[12].

Finally, basing on commonalities among companies that have received a return from their investment in data analytics for their supply chain, at the end of the survey, three factors have been identified: the develop of a strong company-wide analytics strategy, the integration of data analytics in operations for improving decision making process within the whole company, and the construction of a team of people with analytics skills and knowledge of the business. These three factors come to the impacted much more in the results that big data analytics ensures.

To conclude, data analytics offers the opportunity to improve overall operating and financial performance. Obviously, their exploitation implies a sizeable investment and it must be consistent with the company's overall data analytics strategy, but select-

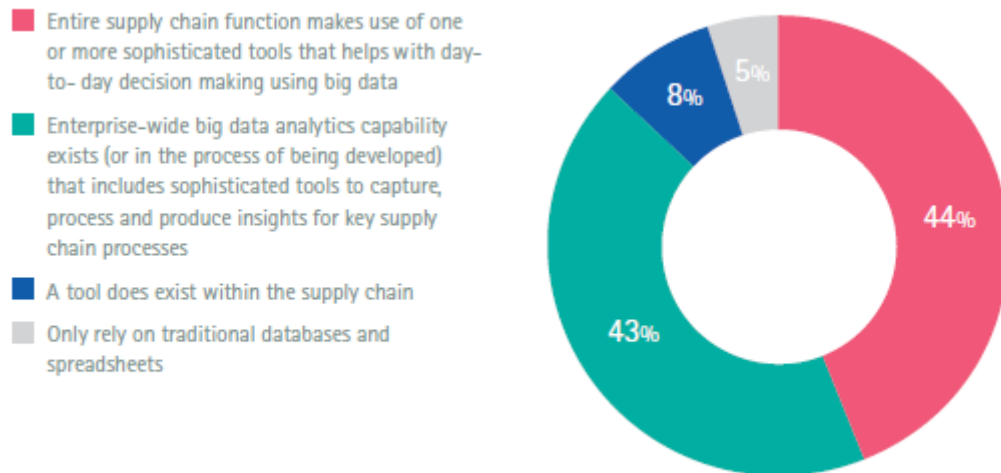


Figure 2.8: How tools and technology support the use of big data analytics[1].

ing the right approach to deploying and scaling a data analytics capability, it will generate a significant business value.

Chapter 3

Dataset review

3.1 Data gathering

The analyzed dataset was collected between April 2018 and July 2019 and includes 2294 distinct items of the manufacturing process analyzed.

It is exploited for other activities within the production control and material programming system, in order to ensure a supply chain visibility. Specifically, data feed a platform that helps to monitor step by step operations and generally supply chain events to avoid interruptions of industrial production due to a lack of raw materials or semi-finished products. These data come from the information system that is used at any time. In this way there is a near real time data exploitation.

Despite being a large dataset, it represents only a limited amount of data used in the entire platform and it is mainly focused on the inventory management domain. It is repurposed for this study.

It consists of three parts, each for one dimensions of analysis: the first one refers to **stock level**, the second one to the **delivered quantity**, i.e. the inbound material received from suppliers, and finally the third one to the **gross requirements**, i.e. the daily products demand from the production plant. Each part is derived from six CSV files, aggregated in a single file through a reading loop in Knime (Figure 3.1). That preliminary process has

been done because of the high dimensions of the data to employ. Indeed, they have been extracted from the source system into six separated pieces, so as to avoid a crash and slowdowns in possible simultaneous activities.



Figure 3.1: Workflow to read multiple Excel files and append them in a single file.

In addition, there is a fourth smaller file which provides information with a more descriptive nature, such as product names and planning phase usage.

A detailed description for each component is offered by the next sections.

3.1.1 Stock

The Stock dataset is the simplest to manage, among the three parts correlated with the critical dimensions of analysis. It gives daily records of the stock during the observation period. Each row represents the instantaneous stock level of a specific date considering the operations made during the preceding day. The data dictionary is as follows:

Attribute	Description
item_code	unique identifier of the item
elab_date_time	reference date of the item stock value in text string format
actual_stock	amount of the same item in the warehouse

Table 3.1: Stock data dictionary.

The peculiarity of this dataset concerns the presence of negative values for the stock. They derive from the way with which materials in inbound are recorded in the corporate management

system. Although it is accounted as an absence, the material is already physically located in the warehouse and its physical movements enters in the system at a later stage[24]. Once the material receipt have been posted, the book inventory balance is no longer negative and coincides with the physical stock. This happens when goods issues are entered before their receipts for organizational reasons. From a financial perspective, the stock is valued at cost value, but negative stock do not have cost implications, so no value for the negative stock must be adopted: in other words the inventory cost doesn't reduce with this negative value.

3.1.2 Delivered Products

The delivered products dataset allows to keep track of deliveries from suppliers occurred during the observation period. It consists of records that indicates the date when quantities entered storage and how much material arrived. Obviously, if nothing was delivered, the delivered value is null for the specific date. The data dictionary is reported in Table 3.2 .

Attribute	Description
item_code	unique identifier of the item
date_from	delivered date in text string format
elab_date_time	elaboration date of the delivered in text string format
delivered_qty	amount delivered

Table 3.2: Delivered data dictionary.

There are two date fields because of the recording mechanism. In fact, although the delivered occurred in a specific day, it is processed the following day by the system. Therefore, when material arrives, the event is not processed immediately and, as a result, in system does not turn out nothing until the following day, when the delivered_qty field becomes equal to the quantity entered the warehouse. Moreover, this record is processed till next Sunday, if it is possible. For this reason, each delivered was recorded many

times: the delivered date is always the same, the processing date is the only one that change. In Table 3.3 can be seen an example of the mechanism described.

item_code	date_from	elab_date_time	delivered_qty
Item X	2019/05/22	2019/05/22	0
Item X	2019/05/22	2019/05/23	250000
Item X	2019/05/22	2019/05/24	250000
Item X	2019/05/22	2019/05/25	250000
Item X	2019/05/22	2019/05/26	250000

Table 3.3: Delivered processing method.

3.1.3 Gross requirements

The Gross requirements dataset is the biggest and the hardest to handle, due to the complexity of the requirement forecasting process already in place in the company. It provides when and how much item quantities is required for the production process. Required quantities are available only in dates a requirement is scheduled for. This makes the time series incomplete from a date standpoint as it reflects the production calendar in terms of working days. The inclusion of rows concerning days without a demand for production, characterized by a null value, would further complicate the dates management and manipulation.

The dataset is a collection of forecasts that include the value of effective required quantity for the production: the final information are given by the actual value. There are two date-time type attribute, as in the case of the previous dataset described: the first represents the day of the requirement, while the second is the date when forecast/actual values are processed. Particularly, elaborations take into account four weeks on a daily basis (the first week is the current week) and ten weeks aggregated on a weekly basis. Specifically, the data dictionary is shown in Table 3.4.

Attribute	Description
item_code	unique identifier of the item
date_from	gross requirement date in text string format
elab_date_time	elaboration date of (effective/forecasted) requirements
gross_req_qty	(effective/forecasted) amount required for the production

Table 3.4: Gross requirements data dictionary.

In this complicated framework, the actual value corresponding to a specific date is the one that has been processed the day before for the next day. In Figure 3.2, a case of actual value identification is shown. It is taken from the pivot exploited to analyze the dataset.

3.1.4 Items descriptions

Unlike other cases, the item descriptions dataset is not a daily register of the item quantity either received or required for the production process or remained in stock. It enriches and supplements information by providing a description of every single item involved in the study.

The dataset integrates what is available with these fields:

Attribute	Description
item_code	unique identifier of the item
item_descr	text string which describe extensively the item
std_cost	standard cost of a single unit of item
mrp_group	material resource planning group.

Table 3.5: Part description data dictionary.

This dataset is relevant because it offers the opportunity to run a clearer analysis because it is easier to carry out assessments and study specific behaviour by knowing what exactly we're dealing with rather than working with unknown text string. It gives also two interesting measures, `std_cost` and `mrp_group`, which will be useful for the study: the former represents the cost of a single item, while the latter provides the planning frequency of require-

sum gross_req_qty	
	2019-03-26
Etichette di riga	
2019-03-27	63198
2019-03-28	62092
2019-03-29	66792
2019-03-30	66362
2019-04-01	66465
2019-04-02	66713
2019-04-03	66721
2019-04-04	66699
2019-04-05	66661
2019-04-06	66649
2019-04-08	68420
2019-04-09	68307
2019-04-10	68398
2019-04-11	68160
2019-04-12	68384
2019-04-13	68178
2019-04-15	68939
2019-04-16	68698
2019-04-17	68082
2019-04-18	67728
2019-04-19	48970
2019-04-22	273030
2019-04-23	
2019-04-24	
2019-04-26	
2019-04-27	
2019-04-29	341888
2019-04-30	
2019-05-02	
2019-05-03	

Figure 3.2: Example of the actual gross requirement quantity.

ments, production and purchase orders. The value of `mrp_group` can be 'daily', 'weekly' or 'obs'. The last case refers to obsolete items for which there isn't obviously a planning

Data integration has been made by exploiting the join operator. In view of the three separated reading of other datasets, that operation has been repeated three times.

In the end, Figure 3.3 shows how the starting point of the study process has been implemented in Knime: in the upper section of the picture there is the workflow related to the extraction, reading and integration process of stock dataset, while in the lower part

there is the one for delivered and requirements datasets, whose workflow is the same.

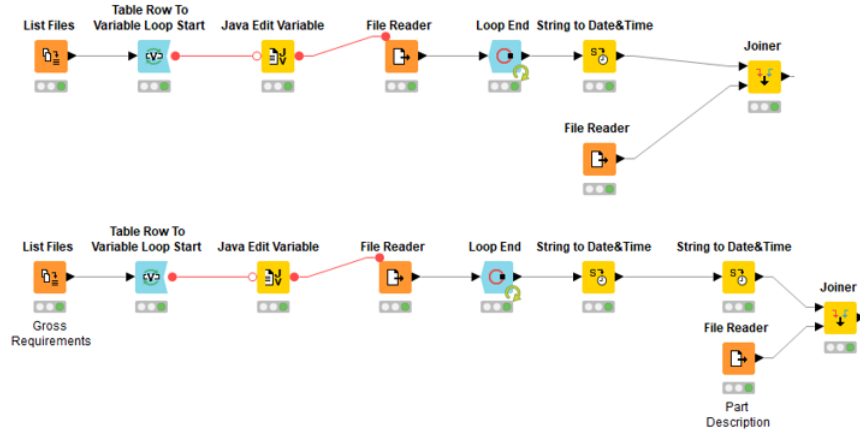


Figure 3.3: Data extraction, reading and integration pipelines

3.2 Main measures statistics

As presented above, the key measures available as input of this study are Stock Quantity over time, Gross Requirements Quantity over time, for different forecasting windows, Delivered Products over time. In the following section an overview of the key statistics about these measures and some graphic analysis will be presented, to start deepening the core dynamics of the studied case.

3.2.1 Gross Requirements Statistics

When analyzing the gross requirements data, it has to be understood that the dataset is made of thousands of products with different pattern. The total quantity of gross requirements over time is clearly showing some gaps (very low production demand), connected with bank holidays or plant halts (see Fig. 3.4).

The data statistics related to daily trend of the total requirements quantity measure show a mean and a median close to 2

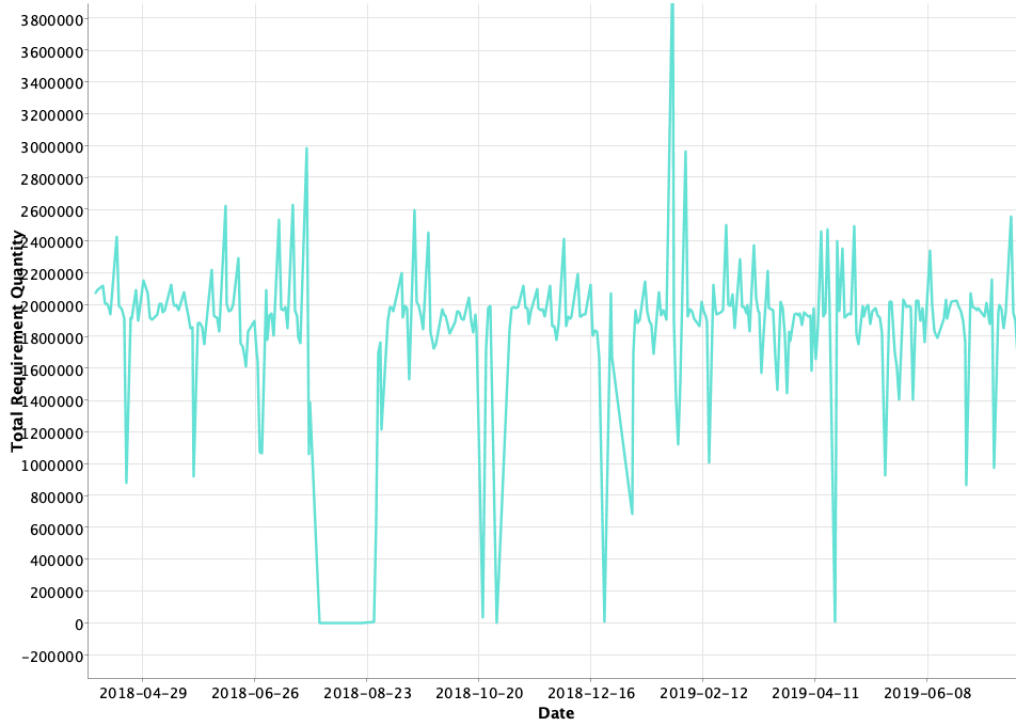


Figure 3.4: Total Requirements quantity trend

millions pieces, with a standard deviation around 400000 units, as per the following figure (Fig.3.5).

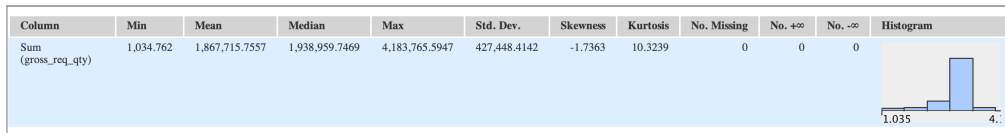


Figure 3.5: Total Requirements Statistics

In order to better understand the underlying dynamics of the different products, it is possible to look at the quartile distribution of the yearly means of each product. Each quartile would then represent a different class of product average daily quantity required: low (0-25%), mid-low (25-50%), mid-high (50-75%) and high (75-100%) product requirement. This is depicted in the Figure 3.6.

For each quartile a clearer representation is offered by the key statistics measures.

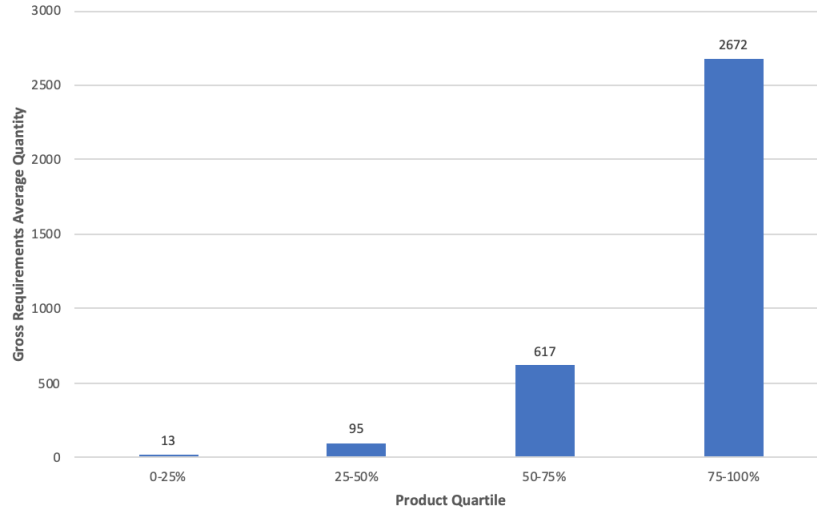


Figure 3.6: Requirement Quantities Quartiles

The first quartile (75-100%), the one with high daily average product requirements, is reported in the Figures 3.7 and 3.8.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Sum (gross_req_qty)	309	1,500,005.5692	1,559,081.96	3,365,752.089	342,266.4257	-1.7779	10.5501	0	0	0	

Figure 3.7: Requirements Statistics of the 1st quartile - high

The second quartile (50-75%), the one with medium-high daily average product requirements, is reported in the Figures 3.9 and 3.10.

The third quartile (25-50%), the one with medium-low daily average product requirements, is reported in the Figures 3.11 and 3.12.

The forth and last quartile (0-25%), the one with low daily average product requirements, is reported in the Figures 3.13 and 3.14.

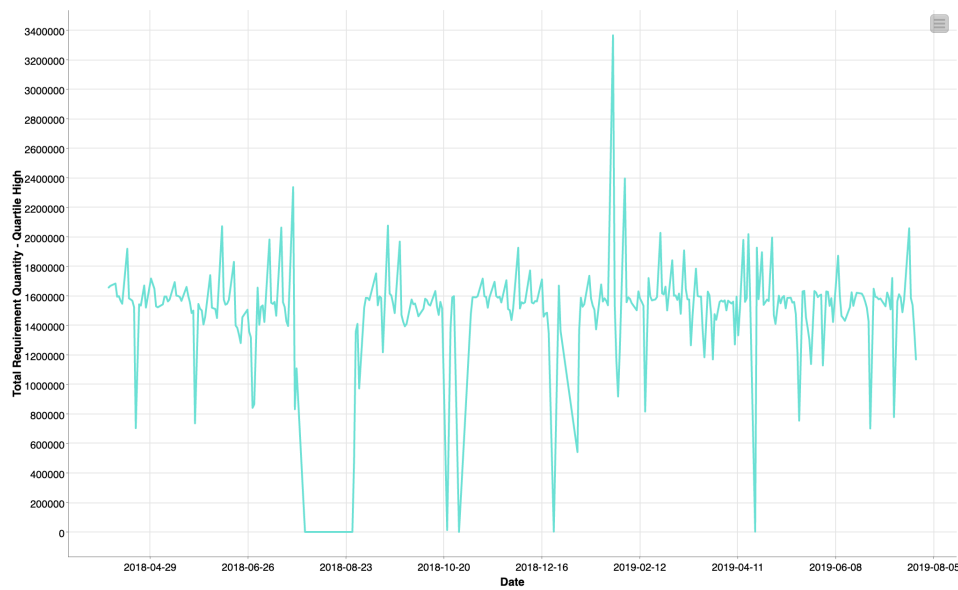


Figure 3.8: Requirements Trend of the products in the 1st quartile - high

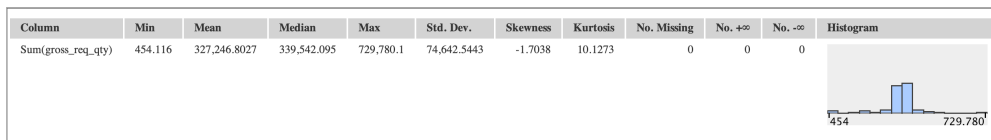


Figure 3.9: Requirements Statistics of the 2nd quartile - medium high

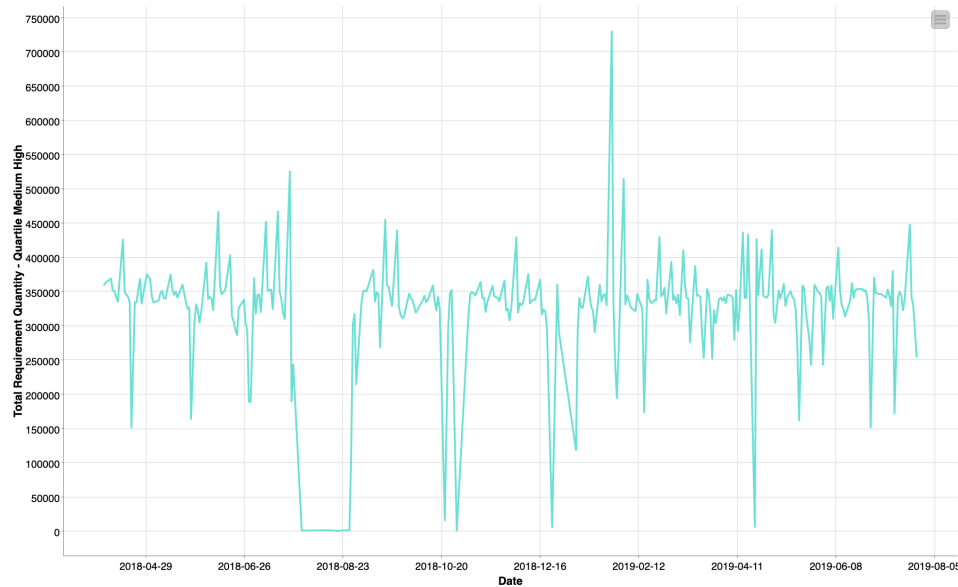


Figure 3.10: Requirements Trend of the products in the 2nd quartile - medium high

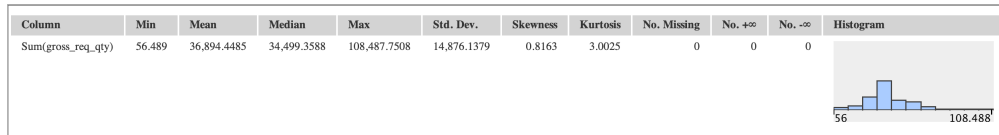


Figure 3.11: Requirements Statistics of the 3rd quartile - medium low

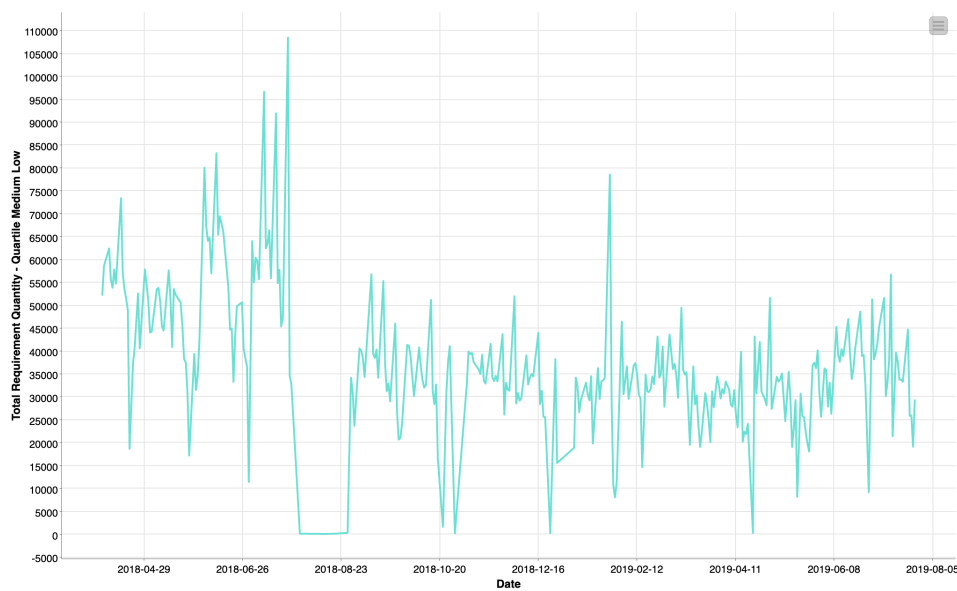


Figure 3.12: Requirements Trend of the products in the 3rd quartile - medium low

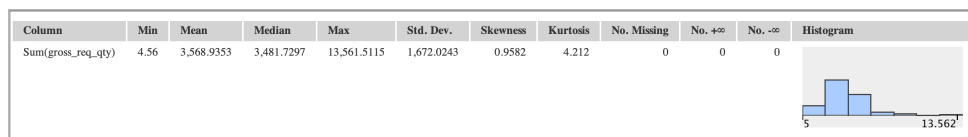


Figure 3.13: Requirements Statistics of the 4th quartile - low

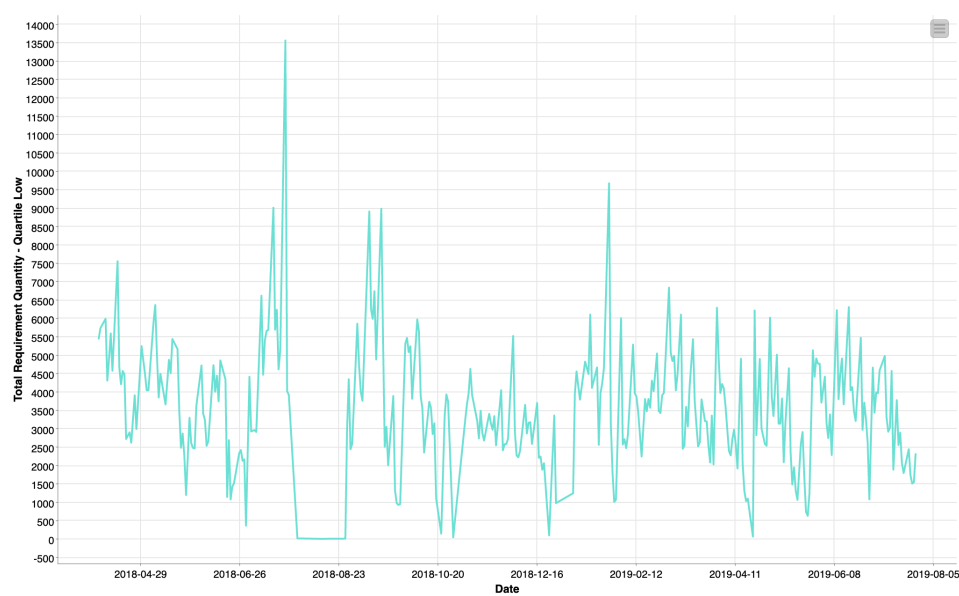


Figure 3.14: Requirements Trend of the products in the 4th quartile - low

3.2.2 Stock Quantity Statistics

Similarly to gross requirements, the stock dataset is made of thousands of products, however in this case each data point represents a punctual daily recording of stock level for each specific product. This means the data cannot be summed across time periods, but must be averaged.

The original stock data as extracted from the Company ERP had a deep negative spike on January 2nd and 3rd, 2019 for many products, very likely due to a glitch (Figure 3.15).

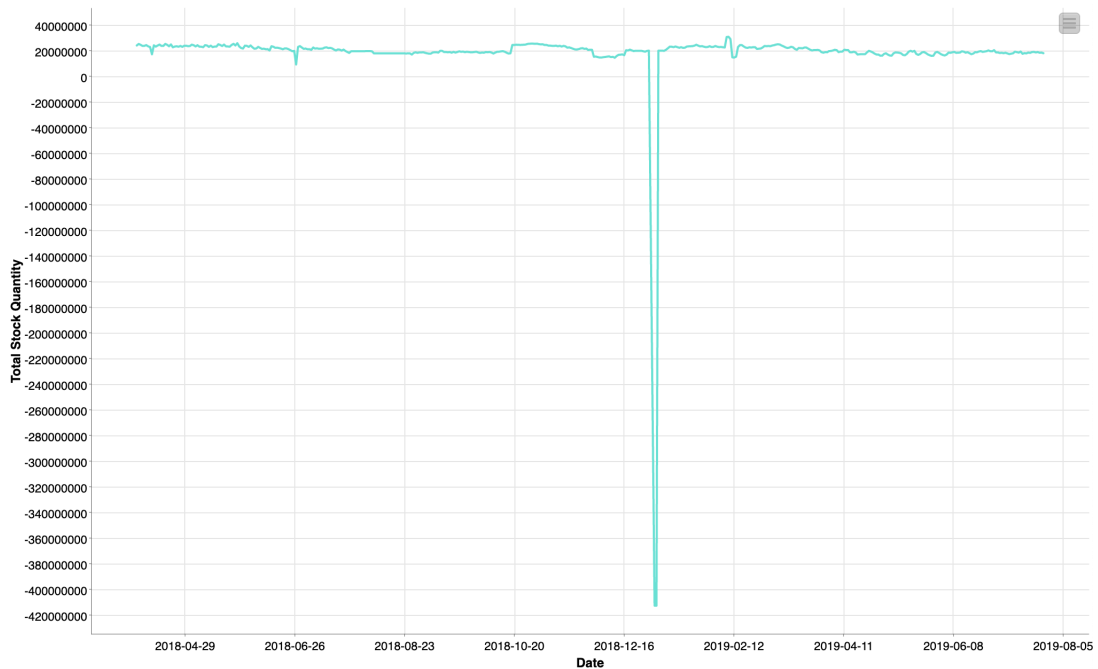


Figure 3.15: Total Stock quantity trend as per initial data extraction

Given these data points were completely outside any confidence level, they have been removed and replaced with previous day data. Similarly, the original stock dataset contained sample products with no business purpose. These have been removed, too, in order to focus only on business relevant data.

The confirmation that these values are outliers has also been given by a statistical approach. Indeed, in order to identify potential anomalies, for each item has been computed the mean and

the standard deviation along the analysis time interval. Assuming the data is normally distributed, if a stock value in a specific date finishes in the tail of the bell curve typical of the normal distribution, that means value is outside the range from -3σ to $+3\sigma$, it can be considered as an outlier. Overall, 1433 of 2294 distinct items are affected by this error on January 2nd and 3rd. As said, the values has been replaced by those of the preceding day which is also the same as on the following day, January 4th. This is a period of holidays and, as a consequence, a production stop period and of plants closure, thus, it can be assumed that these values are derivatives of bad reporting by the ERP system that isn't kept under control in those days.

The overall cleaned stock trend is reported in Figure 3.16.

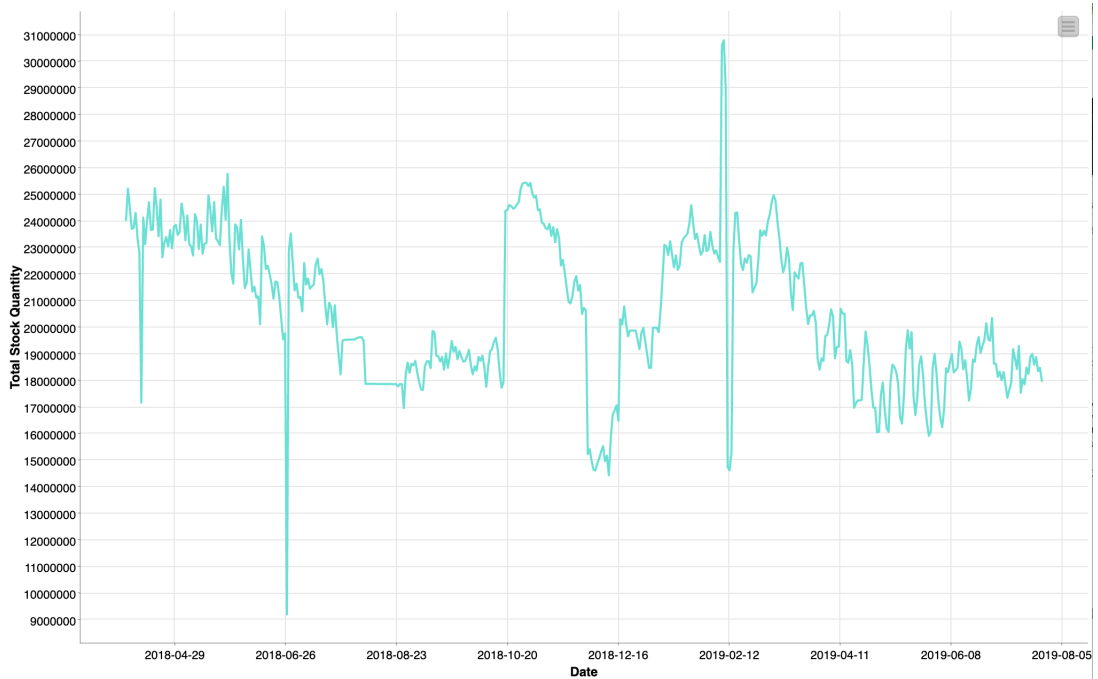


Figure 3.16: Total Stock quantity trend after data cleaning

The data statistics related to daily level of the stock measure show a mean and a median close to 20 millions pieces, with a standard deviation around 2.8 millions units, as per the following figure (Fig.3.17).


Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Sum (actual_stock)	9,189,433.4841	20,528,794.5236	20,043,117.9234	30,786,956.292	2,796,476.0809	0.089	0.1222	0	0	0	

Figure 3.17: Total Stock quantity Statistics

3.2.3 Delivered Quantity Statistics

Last measure described in this section is Delivered Quantity, i.e. the trend of products delivered to the warehouse over time. Similarly to gross requirements, each data point represents the quantity received that specific day, therefore the sum of all data points matches the yearly provisioning of each product. The pattern of this data is much more dynamic and represents well the overall provisioning phenomenon, as shown in Figure 3.18.

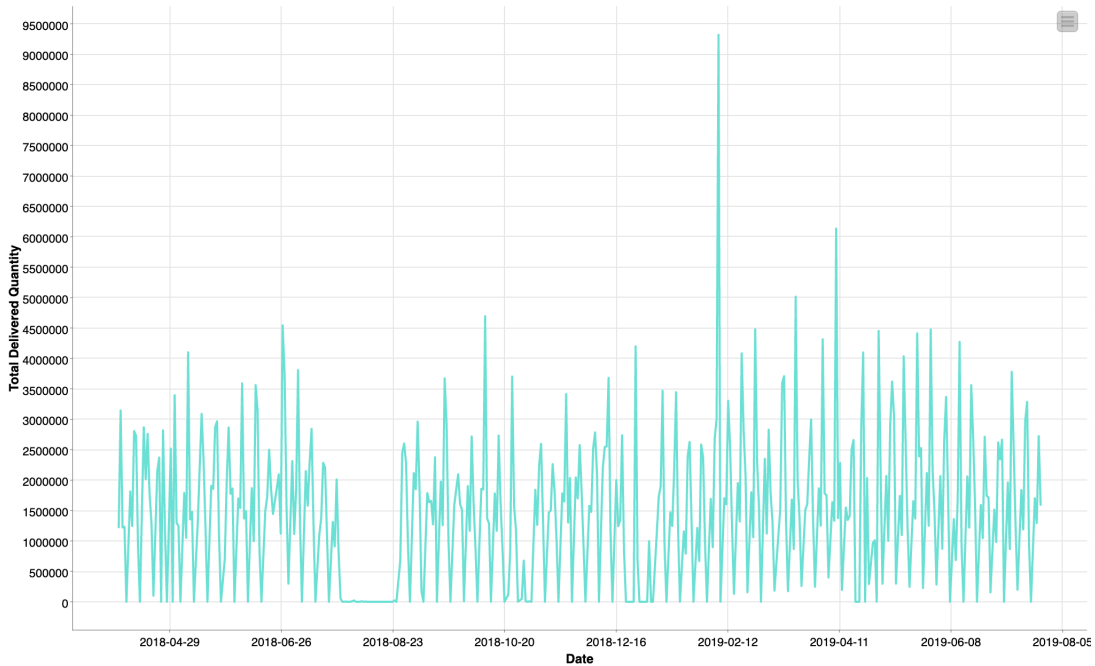


Figure 3.18: Total Delivered quantity trend

The data statistics related to daily trend of the total delivered quantity measure show a mean and a median close to 1.5 millions pieces, with a standard deviation around 1.2 millions units, as per the following figure (Fig.3.19).

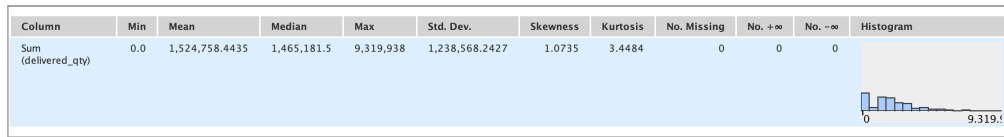


Figure 3.19: Total Delivered quantity Statistics

As done before, we can look at the quartile distribution of the underlying products to observe what trends and behaviors are contributing to the total pattern. Here again each quartile represents a different class of products daily average quantity delivered: low (0-25%), mid-low (25-50%), mid-high (50-75%) and high (75-100%) product delivery. A clear representation can be seen in Figure 3.20

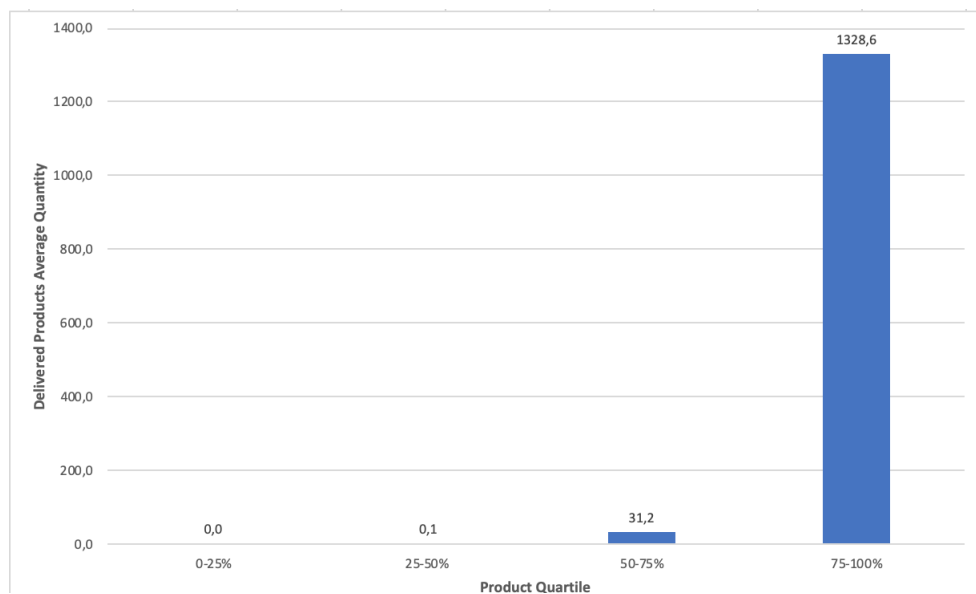


Figure 3.20: Delivered quantities Quartiles

For each quartile a clearer representation of the aggregated delivered quantity is offered by the key statistics measures.

The first quartile (75-100%), the one with high daily average product delivery, is reported in the Figures 3.21 and 3.22.

The second quartile (50-75%), the one with medium-high daily average delivered products, is reported in the Figures 3.23 and 3.24.

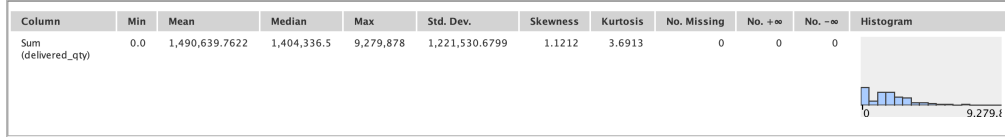


Figure 3.21: Delivered Products Statistics of the 1st quartile - high

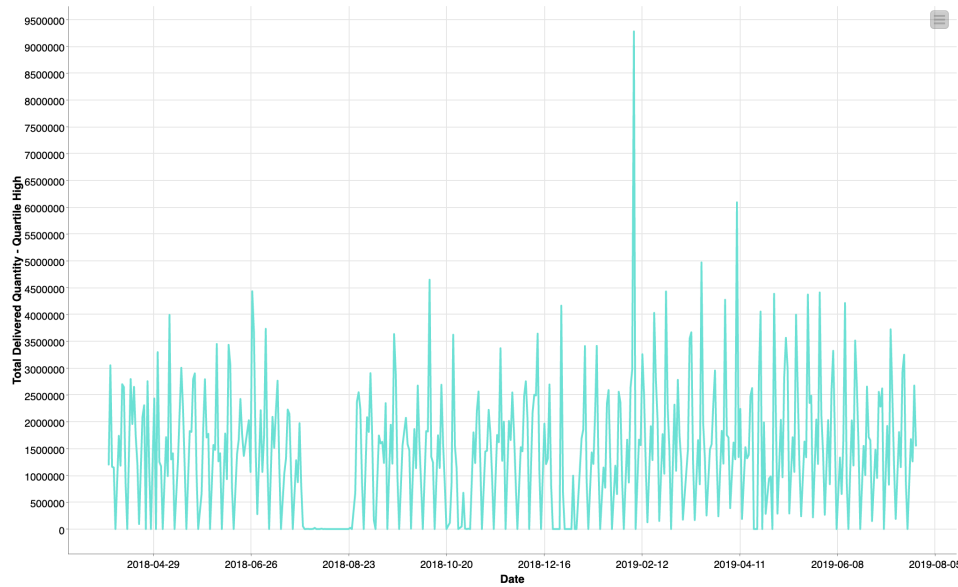


Figure 3.22: Delivered Trend of the products in the 1st quartile - high

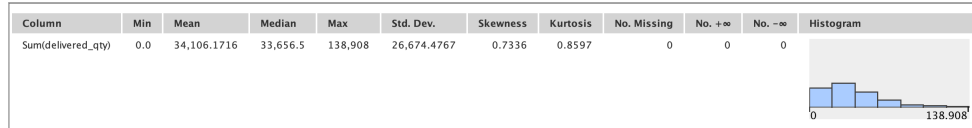


Figure 3.23: Delivered Products Statistics of the 2nd quartile - medium high

The third quartile (25-50%), the one with medium-low daily average product delivery, is reported in the Figures 3.25 and 3.26.

The fourth and last quartile (0-25%) in the case of delivered quantity, contains products with a null delivery (see Figure 3.27). Interestingly, 71% of these products are obsolete, as per the descriptive dataset tagging. Moreover, when looking at the total Obsolete products, 88% fall under the Delivered Products fourth quartile, as per Figure 3.28).

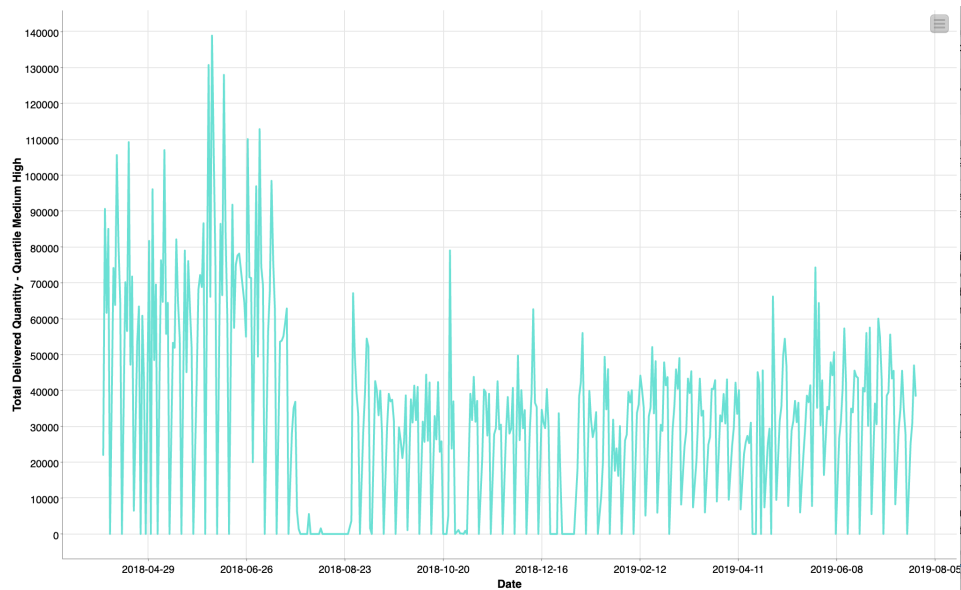


Figure 3.24: Delivered Trend of the products in the 2nd quartile - medium high

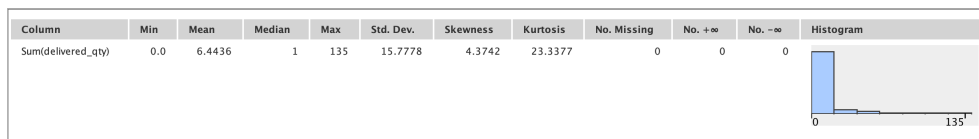


Figure 3.25: Delivered Products Statistics of the 3rd quartile - medium low

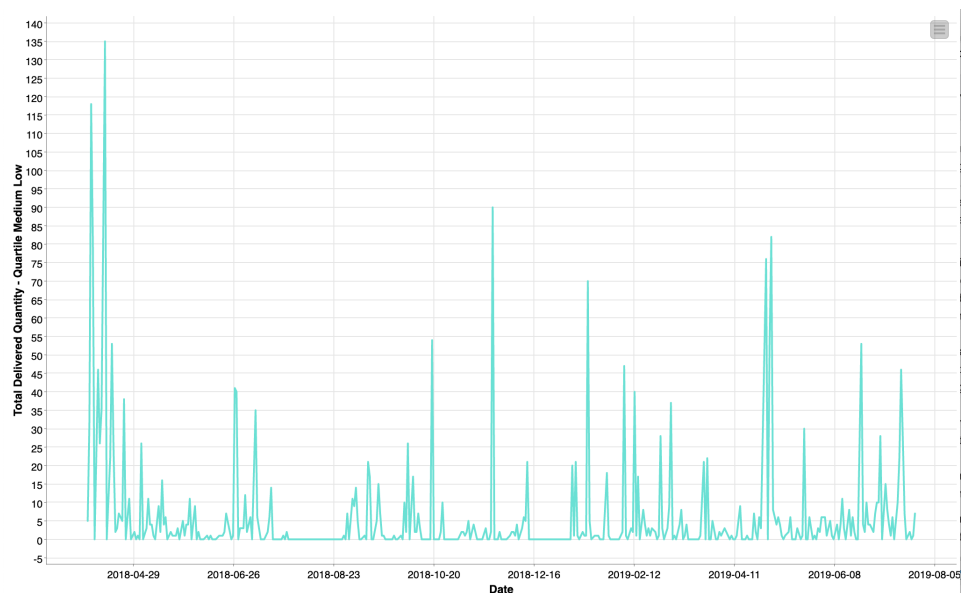


Figure 3.26: Delivered Trend of the products in the 3rd quartile - medium low

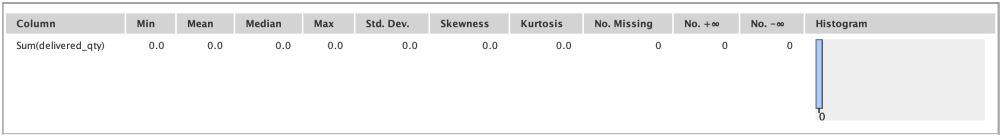


Figure 3.27: Delivered Products Statistics of the 4th quartile - low

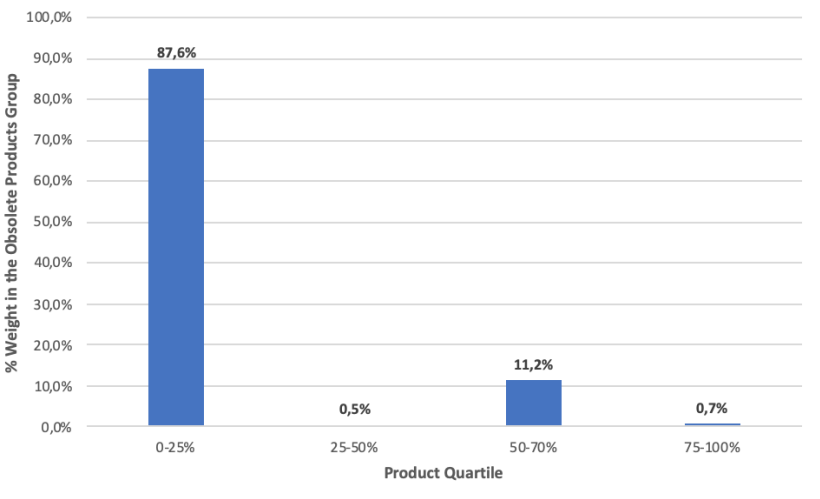


Figure 3.28: Obsolete Products distribution across Delivered Quartiles

Chapter 4

Data Preparation

4.1 Explorative analysis

After the identification and understanding of data sources, the data cleaning has led to more robust data, ready for being exploited. Particularly, in this chapter the raw information are explored further to better understand the business implications of their behavior and dynamics. Relevant insights are transformed in new quantitative elements. This process enriches the original dataset in a way to provide more inputs to models that will be implemented and presented in the following chapters, so that the opportunity to get more interesting and precise results is offered.

Firstly, an ABC analysis has been carried out in order to evaluate items' impact on overall inventory cost and establish a first classification of them. As already mentioned, ABC analysis is a business practice which allows inventory management and control by identifying three items categories. They differ considerably in financial and logistical terms: materials are, indeed, not of equal value and their consumption isn't the same. The resulting consequence is a distinct degree of control that assures a constant production flow, able to satisfy the demands of the customers and, at the same time, to contain the operating costs of the warehouse and the expenses of supplying.

The measures analyzed are stock levels and gross requirements, while delivered material hasn't been taken into account because

it is function of preceding dimensions. Comparing the two ABC analysis we get a 3x3 matrix that enables the identification of service risks or inefficiencies.

The other important objective in exploring the gross requirements data is getting a good understanding of the dynamics over time, in order to understand if patterns or seasonality phenomena are present. Gross requirements represent items' usage during the time horizon subject of study, since they are the demand for feeding production process. Their trends ensure to make observations on how the production is distributed during weeks, months and year. The analysis starts from an overall view of the products portfolio before moving to single item view, looking at their weekly and yearly employment and defining two new measures for each of them.

4.1.1 Cost based ABC analysis

As said in the above paragraph, the analysis has centred around gross requirements and actual stock values quantified on the basis of standard cost. The two dimensions have been treated differently.

Gross requirements constitute a meaningful trend over time, reason why it is important for the analysis to avoid selecting a single day (date) that would have no business meaning. Therefore date from 26-07-2018 to 26-07-2019 has been selected. Furthermore the annual usage value for every item in the dataset has been computed by multiplying the annual gross requirements by the cost per unit.

While, as regards the actual stock, it is an instant value because it always provides a snapshot of the amount in stock of every material and its value in a certain date. Consequently, the sum of values in a given time window isn't meaningful, but to obtain a consistent value that reflects the effective stock's dynamics, the average value of the actual stock for every item in the dataset has been calculated by multiplying the average stock on a cer-

tain period by the cost per unit. Specifically, the analysis have been made looking at three time frame: the last 3 months (from 26-04-2019 to 26-07-2019), 6 months (from 26-01-2019 to 26-07-2019) and 12 months (from 26-06-2018 to 26-07-2019). Eventually, of three classification coming from analysis for each item, the most reliable has been chosen. The selected one, then, has been compared with the classification assigned with gross requirements assessment.

The workflow is implemented for both gross requirements and actual stock values essentially in the same way separately: Figure 4.2 shows the steps performed in Knime.

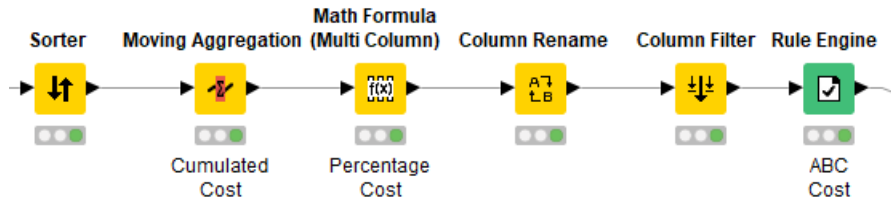


Figure 4.1: Workflow implemented to conduct the cost based ABC analysis.

It starts with data filtering followed by grouping node for defining aggregation values useful for the study. After that there is the mathematical part.

ABC analysis, in fact, requires the application of Pareto's law. Therefore, first thing, items has been arranged in descending order of the usage value computed in the preceding step. Then, a cumulative total of the usage value has been made; hence, it has been possible to compute a cost percentage for each item of the grand total. Finally, a list of rules has been defined:

- cost percentage ≤ 0.80 implies group "A"
- cost percentage > 0.80 and cost percentage ≤ 0.95 implies group "B"
- cost percentage > 0.95 implies group "C"

Each item has received the label of the group for which it has matched the rule.

Before proceeding with classification results assessment, it has been chosen to graphically check beforehand if the logic behind ABC Classification is valid for the warehouse evaluated. The material importance, indeed, should be inversely related with its required quantity and, as a consequence, with the number of items physically presents in warehouse. In this case, only gross requirements are shown, but the complete detailed analysis has been conducted on Knime. Each product annual required quantity has been considered in the analysis, and by going through the workflow described above, has produced the items' percentage of the grand total quantity.

Graphical analysis has led to two graphs: the first graph (Figure 4.2) shows the percentage of each item quantity vs the total decreasingly from the left to the right, while the second graph (Figure 4.3) shows percentage usage value always decreasingly from the left to right. Because of the large size of graph for the high number of items, percentages near to zero have been cut from graphs.

The first two pieces in the above graph have a quantities percentage close to 4% and 3% respectively. They are significant values if it is consider that the inventory includes thousands of distinct items. The same items, however, both have a cost percentage near to 0 so much that they don't appear in the next graph considering that they belongs to its tail. These are low-cost materials but still essential for production process.

Instead, looking to the second graph the first two pieces have a cost percentage close to 10% and 7% respectively, while their requirements quantities percentage are zero: they appear in the tail of the first graph. These are critical items and a right stock needs to be provided because, if they remain unused, the operating and capital costs would be very high and, if they aren't enough, the production process could be blocked.

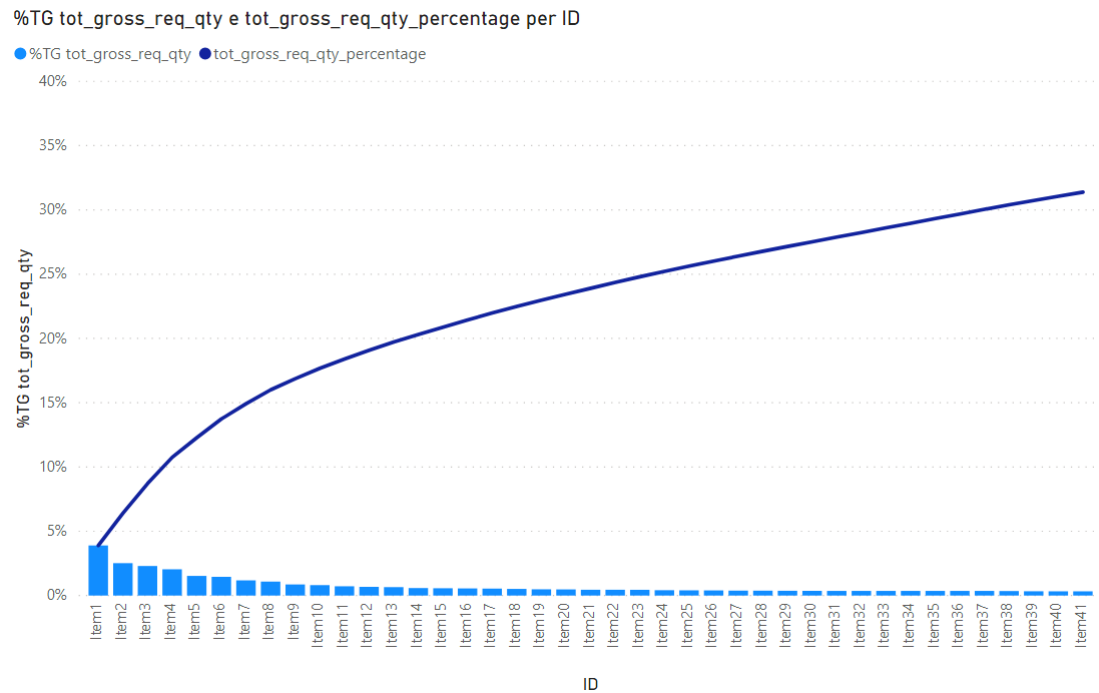


Figure 4.2: Portion of graph that shows the number of items' percentage of their grand total.

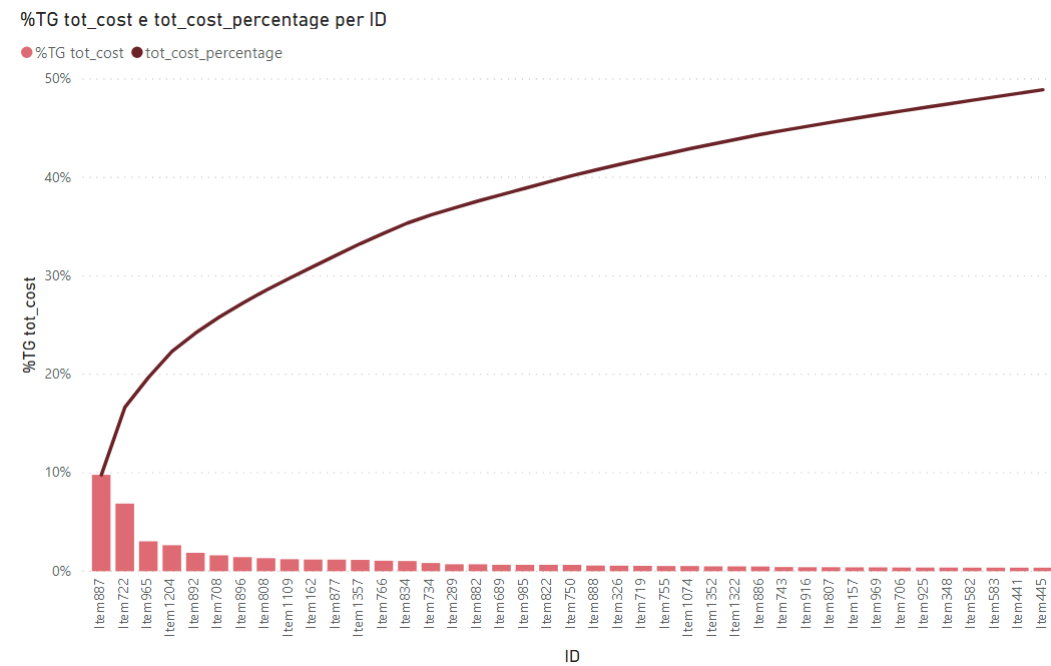


Figure 4.3: Portion of graph that shows the items' cost percentage of their grand total.

Therefore, there must never be an out of stock of items which are critical and have large quantities in the inventory, despite being economic. While expensive material are risky because they constitute tangible fixed assets, but it is always necessary to ensure the right level of service.

Coming back to the main analysis, as regard to actual stock, 3 time periods have been analyzed (Last 3, 6 and 12 months) and the most reliable has resulted the one computed on the last 12 months. The Table 4.1 has been support to the choice. Looking at all possible combinations, there are three very consistent groups which shows that the classification has persisted for almost the entire set of items. Other rows in the table corresponds to slight fluctuations in stock or to destocking decision for reducing the number of items in the inventory: in this latter case there is a gap of two categories during the year.

ABC - 3M	ABC - 6M	ABC - 12M	Count items
C	C	C	1716
B	B	B	385
A	A	A	192
B	C	C	62
B	C	B	61
A	B	B	52
A	B	A	47
C	C	B	37
B	B	A	9
B	B	C	7
C	B	B	5
C	C	A	5
A	A	B	2
C	A	A	2
B	A	A	1
C	B	A	1

Table 4.1: Stock - ABC classification's results over 3, 6 and 12 last months.

At this point, since there is a categorical ABC definition for both gross requirements and stock of each item, it has been possible to build up a comparison matrix (Table 4.2) which counts for

all possible labels combinations the number of items belonging to them.

ABC	Gr.Req - A	Gr.Req - B	Gr.Req - C
Stock - A	152	79	23
Stock - B	32	248	251
Stock - C	48	71	1390

Table 4.2: Stock-Requirements ABC Matrix.

Products in the main diagonal have same ABC class for both gross requirements and stock, which entails they have a stock level which is appropriate to production requirements and vice versa gross requirements justify their amount in the warehouse. The combinations between B and C labels can be considered very similar to the latters. Instead, the combinations between A and C are more delicate because can lead to issues. Particularly, if an item is labelled as A regarding the stock, it means that it is an expensive material with a low level in the inventory. But, if the production requirements for the same item is that of one labelled as C, there is a risk of service because the demand is higher then availability and, thus, to strive it could cause an out of stock. On the other hand, if an item is labelled as A regarding the requirements, it means that it is demanded with low frequency for the production process. But, if it is stocked in large quantities as a C material, there is an inefficiency because a certain part of the stock wouldn't be exploited and, obviously, financially there would be a loss. Although in a lower degree, it is necessary to raise awareness to A, B combinations which could determinate similar problematic situations.

The issue of getting more robust demand and inventory forecasting appears here in order to drive an higher level of service and to avoid inefficiencies. This is what will be attempted to attain with the modelling carried out in the next chapters and which extracts important information from this analysis.

4.1.2 Seasonality and time usage

Before proceeding with the data preparation, model training, and model evaluation, the last step to take has consisted of a study about the production seasonality. The time usage of items, indeed, can differ depending on period which can be characterized by a work peak or by a production reduction or even downtime, thereby affecting increase or decrease production requirements.

The study has been accomplished by exploiting the data visualization on different time scales. Each visualization offers different insight on the time trend of the data. To support the analysis, two weights have been introduced: the first measure is the daily percentages of the weekly requirements grand total, computed by dividing daily requirements by total weekly needs; the second measure is the monthly percentages of the yearly requirements grand total, computed by dividing monthly requirements by total yearly needs. A weekly cyclicity on months instead has not been identified.

From the plot of the monthly time series (Figure 4.4), it can clearly see that the production flow during the year isn't constant. The production takes off from January to reach the highest point in May. After that, a normal production level comes back during June and July, followed by a drop which leads to hit the bottom in August when the production is stopped for summer holidays, made exception for a few days of work concentrated at the end of the month. From September to the end of the year, the production stabilizes again on higher values but on average still smaller than those of the first part of the year. Thus, according to what the graph indicates, most of the production is concentrated in the first half of the year and every holiday results in a production decrease with the summer one more pronounced because of longer duration.

However, it should be clarified that what the graph shows is an average behaviour among the inventory. In fact, there are materials which aren't employed throughout the year. The dataset

includes pieces with an yearly usage that can range to 1 months to 12 months of employments. This is a peculiarity of each item that will be keep in mind in modelling phase.

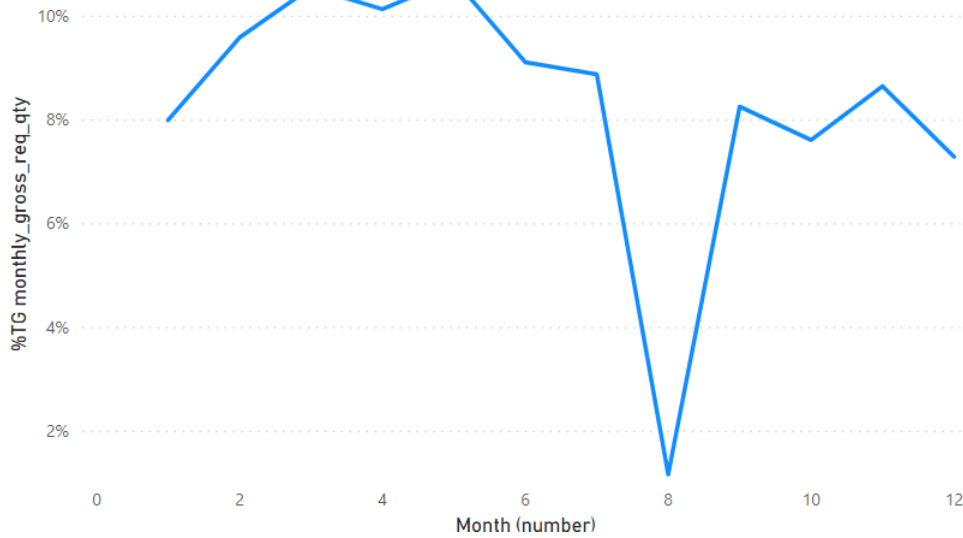


Figure 4.4: Monthly gross requirements trend.

If we switch to the daily scale (Figure 4.5), there is a weekly pattern with the highest point of production of the week on Monday. After this day, the production is almost steady till Saturday which is the lowest point with a reduced production. In this latter case, it is important to clarify that the working weeks of the year not always have a duration of six days, but normally last 5 working days from Monday to Friday, less than that on public holidays.

However, even in this case, the graph illustrates an average trend among the inventory, because items have a different usage during the weeks of the year. For this reason, the number of days of employment has been calculated for each item for each week. Then, this value has been divided by the working duration of the week; in this way, each item has its own weighted average weekly use.

As last activity, the number of weekly usage and yearly usage has been compared in a scatter plot graph to understand how

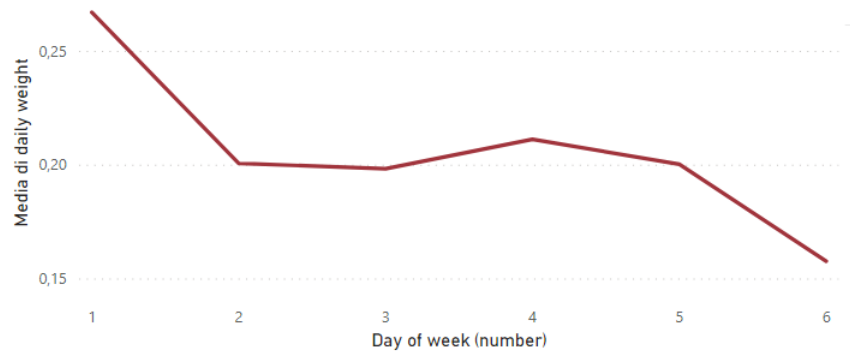


Figure 4.5: Daily gross requirements weight over the week trend in the working weeks

every item behave from a usage standpoint (Figure 4.6).

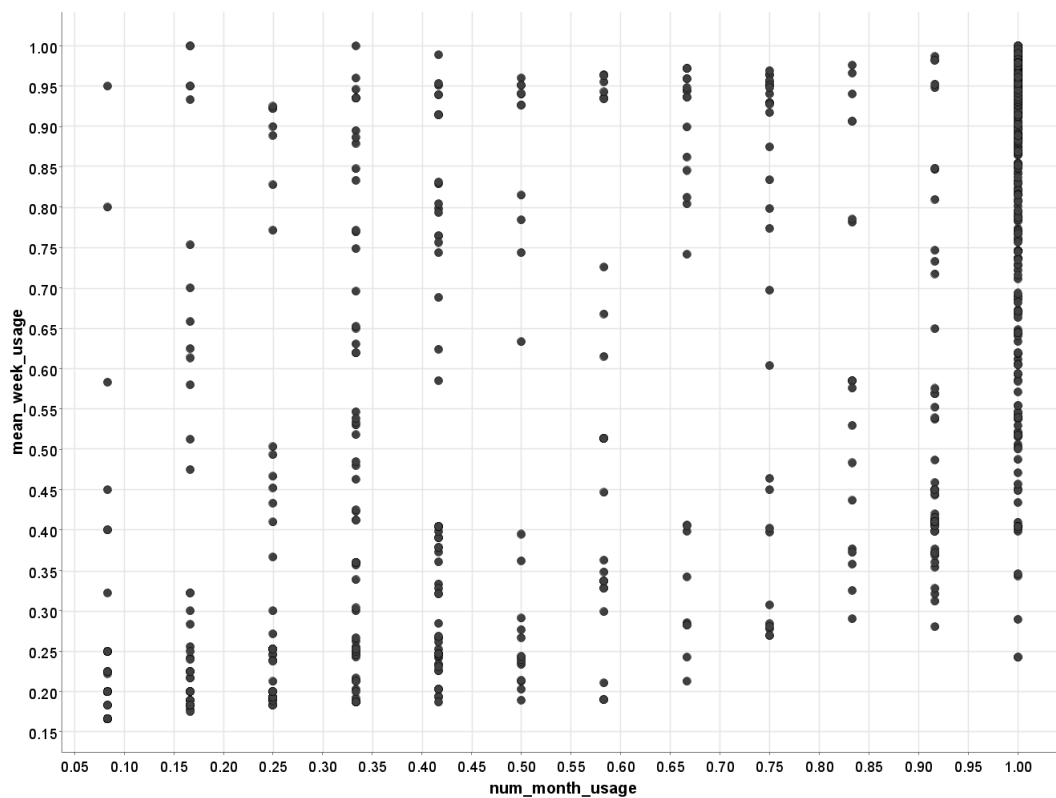


Figure 4.6: Weekly and yearly usage relationship.

While on the right there are materials more utilized during the year with a weekly usage that varies depending on the number of process involving it, the graph shows on the left particular cases.

In the upper part, there are items with an high weekly usage but they are exploited for a low number of months: they are old items which have been decommissioned for instance during the year, or new item still employed for a short time. In the lower left part, there are materials whose employment is reduced both in the weeks and months. They constitute not performing inventory, because there is stock of material inconsistent with its usage. The extreme case of this situation is the obsolescence. Again, hence, at the end of this analysis, as in previous section, the importance of robust demand and inventory forecasting gets back.

In conclusion, analysis results have led to two measures for each item: the average week usage and number of months usage during the year. After looking at all features and characteristics of this data, there are all the essentials to proceed to proceed further with the other steps leading to modelling.

Chapter 5

Modeling

5.1 Cluster analysis

As seen with the ABC classification and seasonality analysis, there are some peculiarities which lead to the inability to treat in the same way all the inventory. It is necessary to recognise similarities and structural differences in terms of costs, quantities and usage between the items, in order to define homogeneous groups so that materials can be managed according to whether they belong to a class or not. The groupings has been made through the clustering, a unsupervised learning method.

For the purpose of this study, the clustering operation is essential. The objective of the dissertation is to get the best data based approach to demand and inventory forecasting. This objective, indeed, will be achieved in a more correct and robust way with the segmentation determined by the clustering operation. There will be a prediction model for each cluster because predictions can't be made in the same way for items with different behavior.

The Clustering has been performed on Knime through the K-Means. Accurate descriptions of the algorithm and of the workflow implemented is in the next sections.

5.1.1 K-Means clustering

K-means clustering is an unsupervised learning technique used for data classification. In unsupervised learning algorithms, no output data is available during the learning process, so automated data exploration techniques are used to identify patterns. K-means uses an iterative refinement method to produce a classification based on a parameter K , defining the desired number of clusters. The algorithm determines the clusters trying to identify a number of centroids (or means) equal to K , that represent the center of each cluster. The first centroids are randomly chosen and all data points are allocated to the cluster corresponding to the nearest centroid according to the Euclidean distance of the points. The K-means then recalculates the centroids value, by averaging the data points belonging to each identified cluster and starts re-assigning the data points to the new centroids.

This steps are repeated as the algorithm iteratively recalculates new means in order to converge to a final clustering of the data points. The algorithm converges when no changes in centroids value occur or no data points change clusters. One of the most challenging aspects of clustering is that the number of desired clusters may not be known a priori. One approach, that will also be applied in the this study, is testing different number of clusters and rank the resulting sum of squared errors. K can be chosen according to the value for which an increase will cause a very small decrease in the error sum, while a decrease will sharply increase the error sum. This is also called “elbow point” and defines the optimal K , therefore the optimal number of clusters.

5.1.2 K-Means implementation

The carrying out of clustering has been arranged in various passages.

First of all, there has been the identification of variables able to effectively differentiate behaviors associated with each item. A

good characterization is made with:

- `avg_actual_stock`: average stock level of the item during the last year;
- `avg_cost_percentage_stock`: cumulative percentage of the grand total inventory cost for the item;
- `tot_gross_req_qty`: amount of item required during the last year;
- `tot_cost_percentage_req`: cumulative percentage of the grand total demand cost for the item;
- `num_month_usage`: item's number of months of employments;
- `mean_week_usage`: item's weighted average weekly usage.

The list of fields selected provides evidence of the importance of the analysis described in the previous chapter. Each variable, substantially, has derived from the elaborations and aggregations made for ABC classification and for seasonality study. Particularly, it has been decided to exploit only the measures (cumulative costs) which have led to the ABC class assignment, rather than the class itself, whatever it was. That was decided because k-means needs the objects to be in numeric format and strings are neither. Obviously, it could be possible to perform a binomial transformation by appending as many columns as possible values defined for the selected column (in these case 6: A,B or C for stock and A,B or C for gross requirements) and by assigning value 1 or 0 according to the class each item belongs to. Nevertheless, there would be the risk that the predictive performance gets degraded because the analysis columns increase. This is the *curse of dimensionality* [10]: it depends on the fact that the distance between the points in a multidimensional spaces tends to become flatter increasing the number of dimensions; in this way the task of the algorithm, which researches meaningful connections between the points, is made more difficult.

In the next step, it has been examined the correlation (Table 5.1) between pairs of attributes for each selected column. According to the matrix, there is a strong positive correlation between actual stock and gross requirements, while there is more moderate positive correlation between item's number of months usage and item's weighted average weekly use and between the two cost components. Moreover, there are slightly negative correlations between costs and their respective quantity references, stock and gross requirements, and between costs and items' employments during the time. The first case has been already highlighted through the Pareto analysis in the previous chapter and may also be influenced by quantity discount from suppliers, a practice which they adopt because larger orders reduce their costs. Therefore, when material is purchased, suppliers give a discount if the order is over a certain size[6]. In the second case it is appropriate to recall the role of economies of scale. They imply a cost advantage obtained due to the production scale, with cost per unit decreasing when production increases: basically, unit costs are generally, but not always, lower for a large plant than for a small one[27]. These are related phenomena because economies of scale provide incentive to order materials in bulk [27] and reduce both production and purchasing costs.

	avg_act_ stock	avg_cost_ perc_stock	tot_gross_ req_qty	tot_cost_ perc_req	num_month_ usage	mean_week_ usage
avg_act_ stock	1	-0.054	0.928	0.046	0.111	0.129
avg_cost_ perc_stock	-0.054	1	-0.068	0.722	-0.204	-0.275
tot_gross_ req_qty	0.928	-0.068	1	-0.014	0.164	0.195
tot_cost_ perc_req	0.046	0.722	-0.014	1	-0.235	-0.297
num_month_ usage	0.111	-0.204	0.164	-0.235	1	0.842
mean_week_ usage	0.129	-0.275	0.195	-0.297	0.842	1

Table 5.1: Correlation Matrix.

Before proceeding with the K-Means implementation, there are other two step to be done. The first one regards the normal-

ization of the values of all numeric columns: the aim is to bring the values of each column on a reduced domain that makes comparison easier. The second one concerns the application of Elbow Method to help finding the appropriate number of clusters, because that number must be specified in input to the K-Means and it's quite hard to estimate alone upfront how many groups there are considering the dimension of analysis. Figure 5.1 exhibits the workflow and its components, executed in Knime.

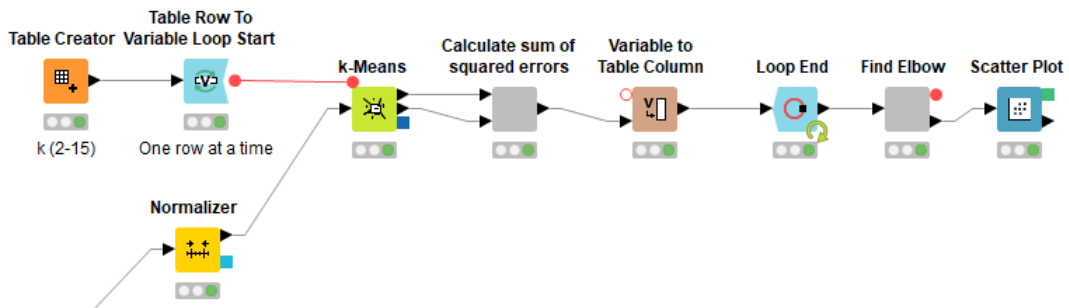


Figure 5.1: Loop to compute Elbow Method.

Data normalised have gone to feed the loop of Elbow Method. Its idea is to run K-Means for a range of values of the number of clusters k . For each k value, the sum of squared errors (SSE) is sequentially computed. It is a measure of quality within cluster: specifically it is the sum of the distances of all points to their respective cluster centers. The SSE value for each k is plotted and the best number of clusters is found where there is an angle in the plot that is a drop in the SSE value.

The workflow has been run over a range from 2 to 15 of k . For each iteration, the value of k has been provided as a flow variable and it controls the setting of the number of cluster of the K-Means, firstly node to be performed. The iteration's SSE computation has started with a join between the two outgoing tables from K-Means: the first contains all items and their relative cluster to which they are assigned, while the second includes cluster's centers. Thus, in each row there are both points and their relevant cluster center. After filtering in only data useful to com-

putation, a 'Java Snippet' code is used to calculate the squared distances for each row. Then the SSE value for the k number of clusters is obtained by the sum of these squared distances, calculated through a "GroupBy" node. This process is the 'Calculate sum of squared errors' component described in Figure 5.2.



Figure 5.2: Calculation of sum of squared errors (SSE) for each iteration.

At the end of the loop, the SSE values for each iteration have been as follows:

Sum Squared Errors	K	Iteration
194.336	2	0
148.898	3	1
101.715	4	2
75.197	5	3
64.748	6	4
60.546	7	5
55.113	8	6
50.64	9	7
49.344	10	8
47.795	11	9
45.063	12	10
42.954	13	11
41.994	14	12
29.876	15	13

Table 5.2: SSE for each iteration.

The best number of clusters has been founded through 'Find Elbow' component and graphically. The result has been the same.

The number of clusters has been determined in the workflow(Figure 5.3), by computing the distances of subsequent SSE

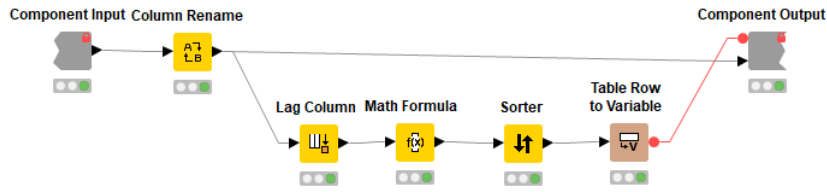


Figure 5.3: Finding of the appropriate number of clusters.

values and sorting the k in order of decreasing distance with the largest one as first (Table 5.3). The best value has turned out to be 4.

SSE	K	Iteration	SSE (- 1)	Delta SSE
101.715	4	2	148.898	47.183
148.898	3	1	194.336	45.438
75.197	5	3	101.715	26.518
29.876	15	13	41.994	12.118
64.748	6	4	75.197	10.448
55.113	8	6	60.546	5.433
50.64	9	7	55.113	4.473
60.546	7	5	64.748	4.202
45.063	12	10	47.795	2.732
42.954	13	11	45.063	2.109
47.795	11	9	49.344	1.549
49.344	10	8	50.64	1.296
41.994	14	12	42.954	0.96
194.336	2	0	/	/

Table 5.3: SSE delta sorting.

Graphically (Figure 5.4), instead, the scatter plot built has provided evidence that, obviously, the SSE decreases as k gets larger: the more number of clusters there are, the smaller the distances between the points and their cluster centers. The optimal value can be seen where the SSE decreases abruptly producing an "elbow" in the graph. The first drop is after $k=3$, so 4 clusters should be taken.

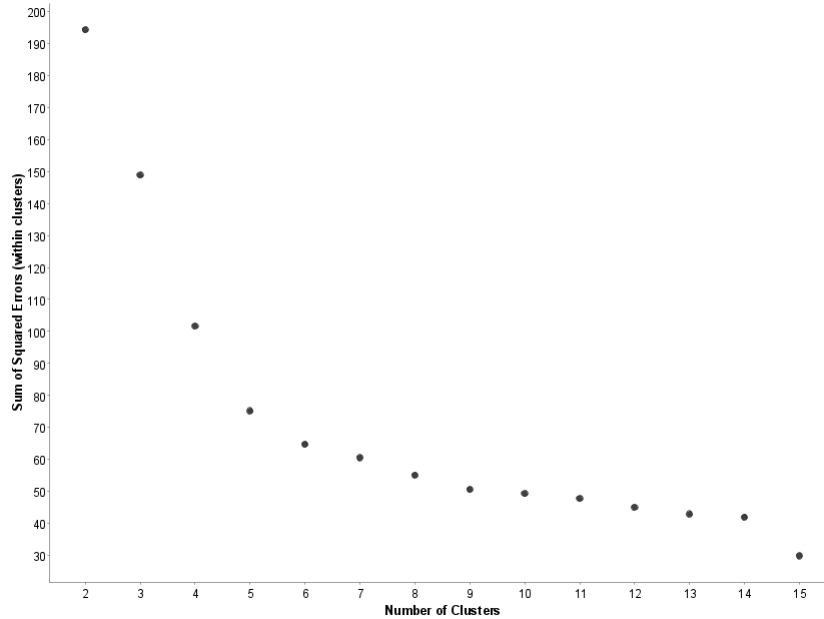


Figure 5.4: Scatter plot of the SSE for all clusterings.

5.1.3 Clustering findings

Once determined the best number of clusters, the K-Means node can be executed.

In order to better understand and interpret the results, data has been denormalized, since there are no more distances to be treated and, consequently, data normalised are no longer essential. At this point, it has been possible to proceed with the analysis in an easier way. Centers of each resulting cluster are listed in Table 5.4.

Cluster	(Mean) avg_act_ stock	(Mean) avg_cost_ perc_stock	(Mean) tot_gross_ req_qty	(Mean) tot_cost_ perc_req	(Mean) num_month_ usage	(Mean) mean_week_ usage
cluster_0	13,257.398	0.968	284,998.817	0.981	0.913	0.98
cluster_1	3,779.476	0.832	171,014.371	0.826	0.96	0.982
cluster_2	571.641	0.982	1,637.795	0.997	0.301	0.296
cluster_3	4,263.947	0.571	173,101.972	0.566	0.965	0.961

Table 5.4: Clusters' centers description.

The scatter plot matrix visualization helps also visualizing the main relationship between key dimensions couples. For a better readability the Figure 5.5 represents 3 out of 6 dimensions, one

per each logical area (volume, value, frequency).

Starting from the cluster's centers data, a detailed description is provided below.

- Cluster 0 includes the items with the highest production requirement quantity, the highest stock and low marginal unit cost. These products have high frequency, both in terms of weekly and yearly usage.
- Cluster 1 includes products with medium average stock level and requirements, mid-low marginal unit cost (slightly higher in ranking versus cluster 0), high weekly and monthly usage.
- Cluster 3 shows stock and requirement quantities very close to Cluster 1. Indeed, stock and gross requirements values belong to the same order of magnitude. Also time usage is very close. The key difference is instead on the cost dimensions, whose center stays in a much more valuable position of the cumulative distribution for both stocks and requirements. In fact, this could be labelled as the "Premium" product cluster.
- Cluster 2 instead is centered around much lower levels of stock and requirements, with a very marginal cost implication, and, more interestingly, very low usage frequency, both from a weekly and a monthly standpoint. This cluster really groups the tail products, which have a minor role in the production cycle or may even be discontinued. On this topic, it is really important to observe that Cluster 2 includes the vast majority of products marked as obsolete in the original MRP classification, even if this was not a parameter used in our unsupervised learning (see details in Figure 5.6 and Figure 5.7). Cluster 2 is then a good predictor for Obsolescence, i.e. products assigned to cluster 2 may be or may become obsolete in near future, so they deserve attention.

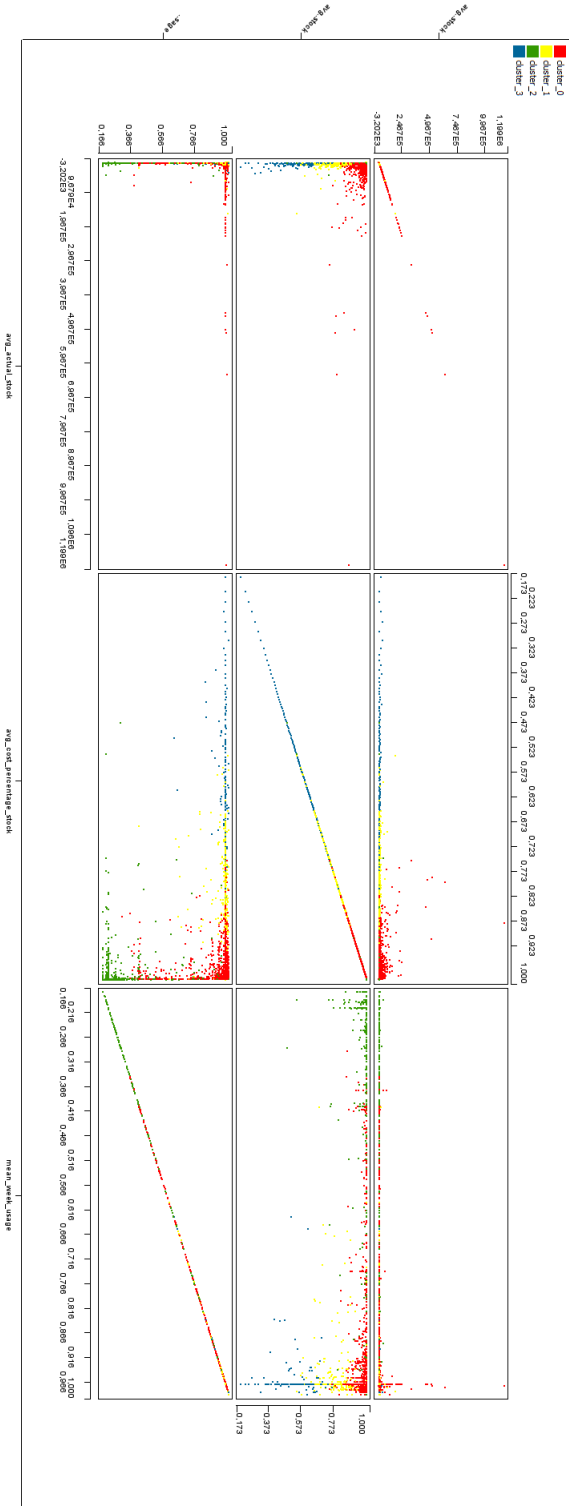


Figure 5.5: Scatter plot matrix with cluster evidence - key dimensions relationship

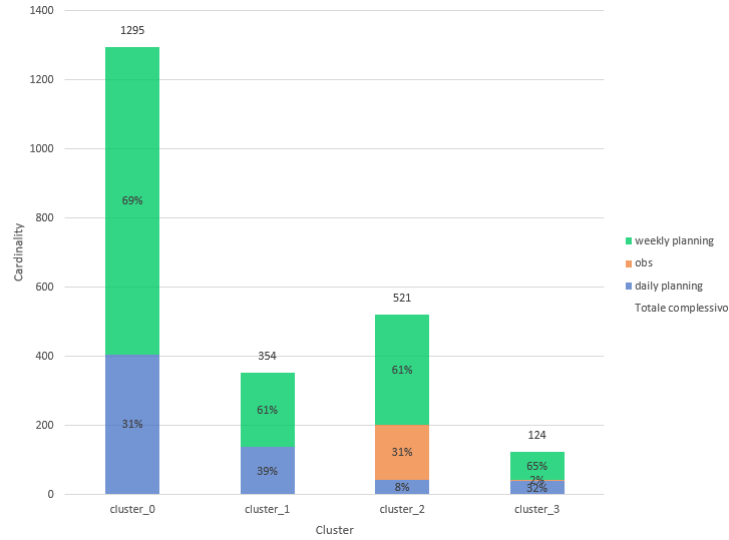


Figure 5.6: Cluster Cardinality and MRP classification comparison

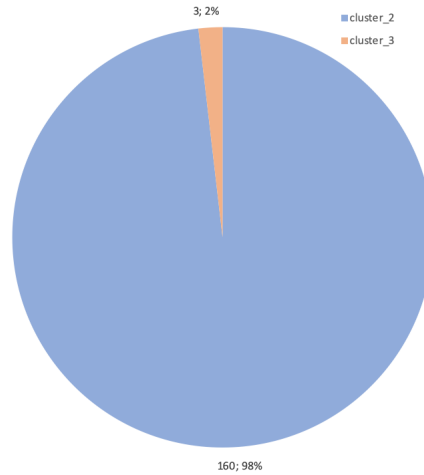


Figure 5.7: Obsolete Products presence in each cluster

5.2 Forecasting

The demand and stock prediction is a time series analysis problem. There are, indeed, time series of numerical values, actual stock and gross requirements per day.

Therefore, the issue requires a supervised learning: it is necessary to build a model which is able to predict the next value given the past N values. More specifically, in this case, firstly,

gross requirements must be predicted; then the obtained values must feed the prediction of items stock value. The idea is that the current stock derives from the stock of the preceding day to whose production requirements have been subtracted.

Within each of the product clusters previously identified, the problem has been dealt by carrying out two alternative learning methods: the Random Forest Regression and the Multilinear Regression. They have been compared and the best one between them has been selected as a solution to the problem. The models will be described carefully with their results in the following sections. However, they share the preliminary workflow related to their implementation, following the approach described above.

5.2.1 Approach to forecast

The dataset has been partitioned into the training set and test set. The split between the two sets has been a split in time: it is reserved the data from April 2018 to June 2019 for the training set and the data of July 2019 for the test set.

At this point, attention has been given to gross requirements whose prediction is the first to be determined. For each value $x(t)$ of a specific item's gross requirements time series, it has been defined the vector $x(t-N), \dots, x(t-2), x(t-1), x(t)$ through a lagging operation (Figure 5.8). The past values $x(t-N), \dots, x(t-2), x(t-1)$ will be the input to the model for the learning phase with the current value $x(t)$ as the target column to train the model.

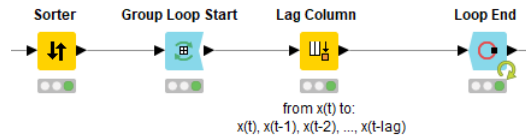


Figure 5.8: Building of items' past N values vector for gross requirements.

The study has experimented separately with two values, corresponding to different vector dimensions: $N=7$ (days) and $N=30$

(days), reflecting the possible influence of weekly or monthly dynamics.

The vector of past values has been built after partitioning the dataset for avoiding data leakage from neighboring values. Moreover, in the test set have been considered days in the interval starting from 01/07/2019 - N days, respectively 23/06/2019 and 31/05/2019 for N=7 and N=30. That was needed in order to avoid building a past values vector having first day records lacking past values, hence affecting the accuracy of forecasts.

elab. date_time	gr _qty	gr _qty(-1)	gr _qty(-2)	gr _qty(-3)	gr _qty(-4)	gr _qty(-5)	gr _qty(-6)	gr _qty(-7)
2019-07-09	2,036	1,978	0	0	2,024	2,008	2,012	2,020
2019-07-10	2,020	2,036	1,978	0	0	2,024	2,008	2,012
2019-07-11	1,934	2,020	2,036	1,978	0	0	2,024	2,008
2019-07-12	2,296	1,934	2,020	2,036	1,978	0	0	2,024
2019-07-13	984	2,296	1,934	2,020	2,036	1,978	0	0
2019-07-14	0	984	2,296	1,934	2,020	2,036	1,978	0
2019-07-15	1,976	0	984	2,296	1,934	2,020	2,036	1,978

Table 5.5: Portion of an items' past 7 values vector.

Once the past values vector has been built, the training model has been defined. The attributes on which the model has been learned are:

- Day of week;
- Month;
- Gross requirements past values vector.

It have been chosen to insert two time measures such as number of the month and the number of weekday because of usage dynamics over time described in the previous chapter.

Then, the model has been tested in order to predict production requirements' values for every day of July 2019. Because the aim is to run the predictions for multiple days after the next one, it has been necessary to loop around the model by feeding the current prediction back into the vector of past values. For this reason, a recursive loop has been implemented.

For each loop iteration, one day of the tested interval has been predicted. At the beginning of the iteration, in fact, the test set

is split every time in two parts: the first one includes the data, corresponding to the current day, to be provided to the Predictor node, while the second one consists of all other data referred to both preceding and following days. The reason for the split is to ensure the possibility to rebuild an 'updated' test set at the end of the iteration.

Specifically, after the forecasting operation, daily prediction becomes the $x(t)$ value of the same day. This is made by acting at column names' level: columns are managed with sorters and transpositions, so that the actual value is replaced with the foretold one. Naturally, before these manipulations, values are stored as Prediction of the day. Finally, recovering all the other days excluded from the analysis, at the end of the iteration the test set is reconstructed by rows concatenation. Before that, the lagging operation's resultant columns are removed from the old data, because a new past values vector must be build with forecasts. Thus, the predicted value becomes the $x(t-1)$ value of the following day and, generally, the $x(t-N)$ value of n -th next day. So, data can be passed back to loop start for the prediction of the next days. Figure 5.9 provides the articulated workflow for the described operations.

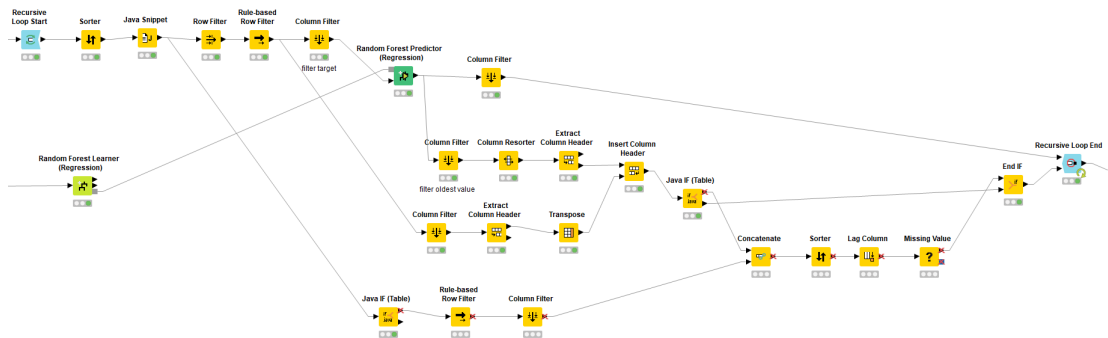


Figure 5.9: Recursive loop carried out to predict multiple days after the next one.

The recursive loop approach has been employed both for the Random Forest Regression and the Multilinear Regression.

Concluded the activities on gross requirements, the focus has shifted on stock values. The same procedure carried out for the gross requirements has been followed, except that, while building stock's past values vector, the lagging process related to t-N values calculation has also been extended to gross requirements and delivered products (Figures 5.10). In this way, for every item there are three past values vectors, one for each analysis dimension, in order to ensure the best possible learning, since stock changes over time in relation to the dynamics of both production forecast and delivery of supplies.

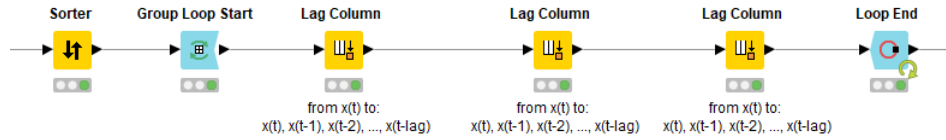


Figure 5.10: Building of items' past N values vector for stock values.

It shall be reinforced that all of these operations have been carried out separately and independently of each other, for each product cluster previously identified, with the aim of identifying the best possible result.

Before presenting the results, in the following 2 paragraphs, a brief description of the two forecast modeling techniques chosen for comparison is presented.

5.2.2 Multilinear Regression

The Multilinear Regression (or multiple linear regression) is an extension of the simple linear regression (or ordinary least-squares regression), using several explanatory variables to predict the outcome of an output variable. This type of regression models the linear relationship between the independent variables and dependent variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik} + \epsilon$$

where, considering $i = k$ observations, we have:

- y_i = dependent variable
- x_i = independent variables
- β_0 = constant term or intercept
- β_k = slope coefficient for each independent variable
- ϵ = residual error term

The multilinear regression is based on few assumptions:

- Dependent variables and independent variable have a linear relationship
- Independent variables are not strongly correlated among themselves
- Residuals are normally distributed
- All observations y_i and regression residuals should be normally distributed

It has to be noted that the R-squared of this model should be carefully evaluated, because it increases with the number of predictors included in the regression, even if these may not be related to the dependent variable.

Beta coefficients of the final equation can be used to interpret the results, while keeping all variables constant.

5.2.3 Random Forest Regression

Among the machine learning techniques, ensemble learning uses a combination of different models to build predictions that are less biased and less data sensitive (i.e. have lower variance). Random forest is an ensemble model using bootstrap aggregation (also known as bagging) as the ensemble method and decision tree as the individual model. In fact Random Forest uses bagging on multiple CART (Classification and Regression Tree) models,

originally having a high variance output, to bring back a forecast that is on average closer to the actual result, granting higher accuracy.

In order to reach this result, the Random Forest algorithm leverages random sampling of training data points while building the decision trees and random subsets of features while diving the nodes. The overall algorithm development can be divided in 4 phases:

- Phase 1 - From the entire training set, n random subsets are selected.
- Phase 2 - n random Decision Trees are trained. Each random subset is used to train one decision tree and the optimal splits for each decision tree are based on a random subset of features, selected from the total features injected in the model.
- Phase 3 - Each individual tree predicts the records in the test set, independently.
- Phase 4 - Final prediction averages the output of the decision trees

The phases 3 and 4 are repeated for each record in the test set. A supportive graphical representation is depicted in Figure 5.11

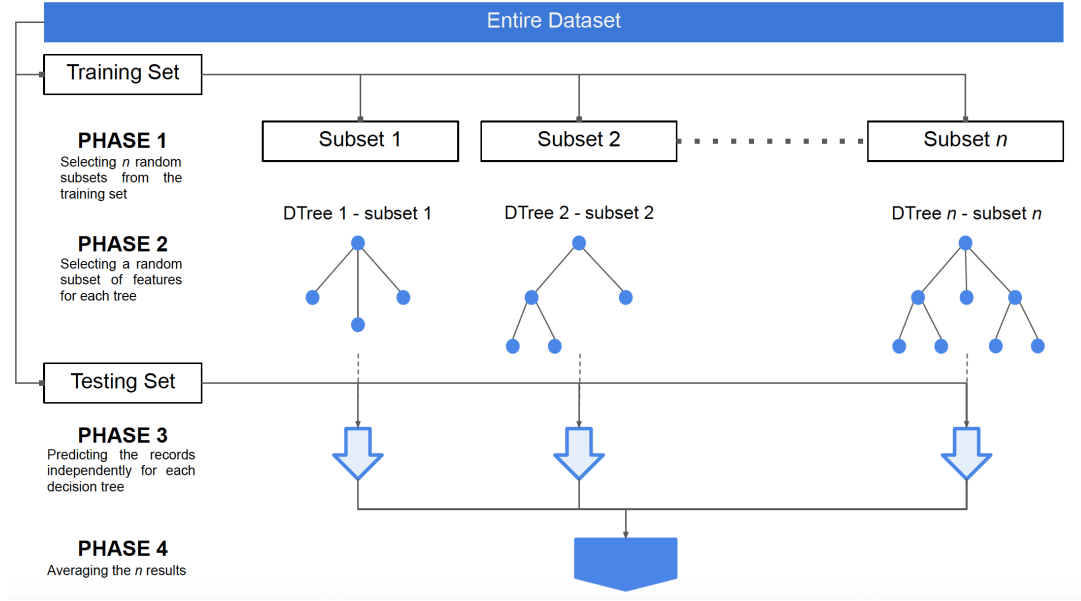


Figure 5.11: Random Forest algorithm schema.

5.2.4 Model evaluations tools

The evaluation of how reliable are the forecasts relies on how the prediction made on the test is close to reality. Performances have been assessed by comparing the target column's values and predicted values determined by the models. There have been two ways of proceeding.

From a mathematical-statistical point of view, two statistics between the compared values have been considered: R^2 , and *Root Mean Squared Error*, (*RMSE*) have been taken into account.

R^2 is defined as:

$$1 - \frac{\sum_{i=1}^N (r_i - p_i)^2}{\sum_{i=1}^N (r_i - \bar{p})^2}$$

where r_i and p_i are respectively the observed outcome and the predicted one, while \bar{p} is the arithmetic mean of the observed outcomes. It provides a performance measure based on the proportion of total variation of outcomes explained by the model. The higher the measure the better the model effectively captures the reality.

RMSE is defined as:

$$\sqrt{\frac{\sum_{i=1}^N (r_i - p_i)^2}{N}}$$

where r_i and p_i are respectively the observed outcome and the predicted one, while N is the number of the rows. It indicates the average level of error of the model: the higher the index value the greater the error made.

From a business point of view, it is more interesting to deal with measures of forecast accuracy closer to the supply chain practitioners. Generally the ratio between forecasted values and actual values or mean absolute percentage error (MAPE) are used to evaluate forecast accuracy. MAPE is defined as follow:

$$M = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

In our specific study case, for both requirements and stocks we will have a Mean Accuracy or MAPE calculated over the 30 forecasted days, as well as a weighted accuracy over the entire month, calculated as absolute value of 1 minus the ratio between the sum of requirements forecast and the sum of requirements actuals across the entire month. The two measures answer different questions, i.e. how close to reality was the daily forecast each day and how close to reality was the entire month forecast, respectively.

The results obtained are presented separately for each cluster identified above and compared to drive the most business meaningful choice.

5.2.5 Cluster 0

Products belonging to Cluster 0 are the highest in number and are characterized by high volume of requirements and stock, low marginal unit cost and high usage frequency. The modeling is

challenging here as there are many different products, but the accuracy comparison among the different model outputs would privilege Random Forest Regression based on 30 previous days datapoints. Interestingly, the stock accuracy measure is better than the requirements' one, despite this latter is fed into the stock forecasting. This demand forecast accuracy risk is however mitigated by the limited cost impact of the products belonging to this cluster.

	REQUIREMENTS				STOCK			
	Random Forest Regression		Multilinear Regression		Random Forest Regression		Multilinear Regression	
CLUSTER_0	7 days	30 days	7 days	30 days	7 days	30 days	7 days	30 days
MAPE	28,78%	28,01%	41,35%	44,16%	4,18%	3,03%	9,88%	5,40%
Accuracy over month	43%	49%	34%	35%	3%	1%	10%	5%

Table 5.6: Accuracy performances of cluster 0 predictions.

The confirmation that Random Forest Regression with a 30 days backward looking feed works better with stock and gross requirements has been given by statistics associated with the model. As can be seen in Table 5.7, it provides the minimum value of $RMSE$ and the maximum value of R^2 for both dimensions.

	Random Forest Regression				Multilinear Regression			
	7 days		30 days		7 days		30 days	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
gross_req_qty	0,93	872,52	0,95	841,53	0,92	1026,15	0,92	869,08
stock	0,95	9685,63	0,95	10635,56	0,95	10380,02	0,96	9543,28

Table 5.7: Statistics of cluster 0 predictions.

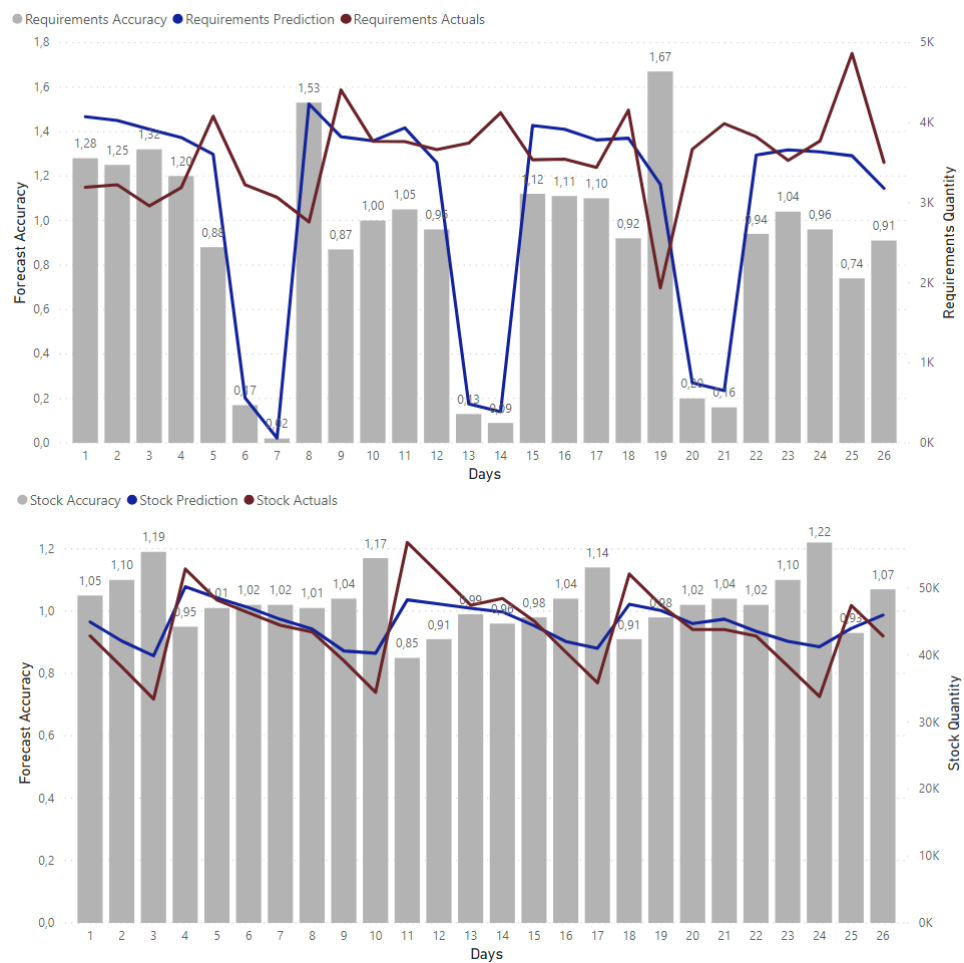


Figure 5.12: Example of results of predictions for Cluster 0 items both for stock and gross requirements.

5.2.6 Cluster 1

Cluster 1 consists of items with a medium production contribution which is reflected in a feeding amount of the same level in the warehouse. These items, employed with an high frequency, affects company's costs in a mid-low way. Looking at results, they appear to lead to the Random Forest Regression based on 30 previous days datapoints as regards stock values: in fact, selecting the Multilinear Regression for stock determines a loss in MAPE value by at least 5 percentage points. Instead, the choice is more complicated in the case of gross requirements values: selecting Multilinear Regression with a 30 days backward looking feed would improve the MAPE value but it would worsen the accuracy value computed over month; the opposite situation would occur if the Random Forest Regression would be chosen. For this reason, both the case of the application of Multilinear Regression for requirements and Random Forest Regression for stock, and the case of the application of the Random Forest Regression only for the two dimensions, have been both tested.

	REQUIREMENTS				STOCK			
	Random Forest Regression		Multilinear Regression		Random Forest Regression		Multilinear Regression	
CLUSTER.1	7 days	30 days	7 days	30 days	7 days	30 days	7 days	30 days
MAPE	15,37%	12,16%	11%	10,88%	5,59%	2,88%	8%	7,81%
Accuracy over month	13%	7%	9%	8%	5%	3%	2%	4%

Table 5.8: Accuracy performances of cluster 1 predictions.

The decoupled models between gross requirements and stock determines results indicated in Table 5.9: comparing them with those obtained by applying Random Forest Regression for stock and gross requirements, they are very close, but there would be an accuracy reduction if we try to align the model on the two sides. Therefore, in view of a better accuracy, it is necessary use a single model for both dimensions.

This has been confirmed by the statistics characterising the model (Table 5.10).

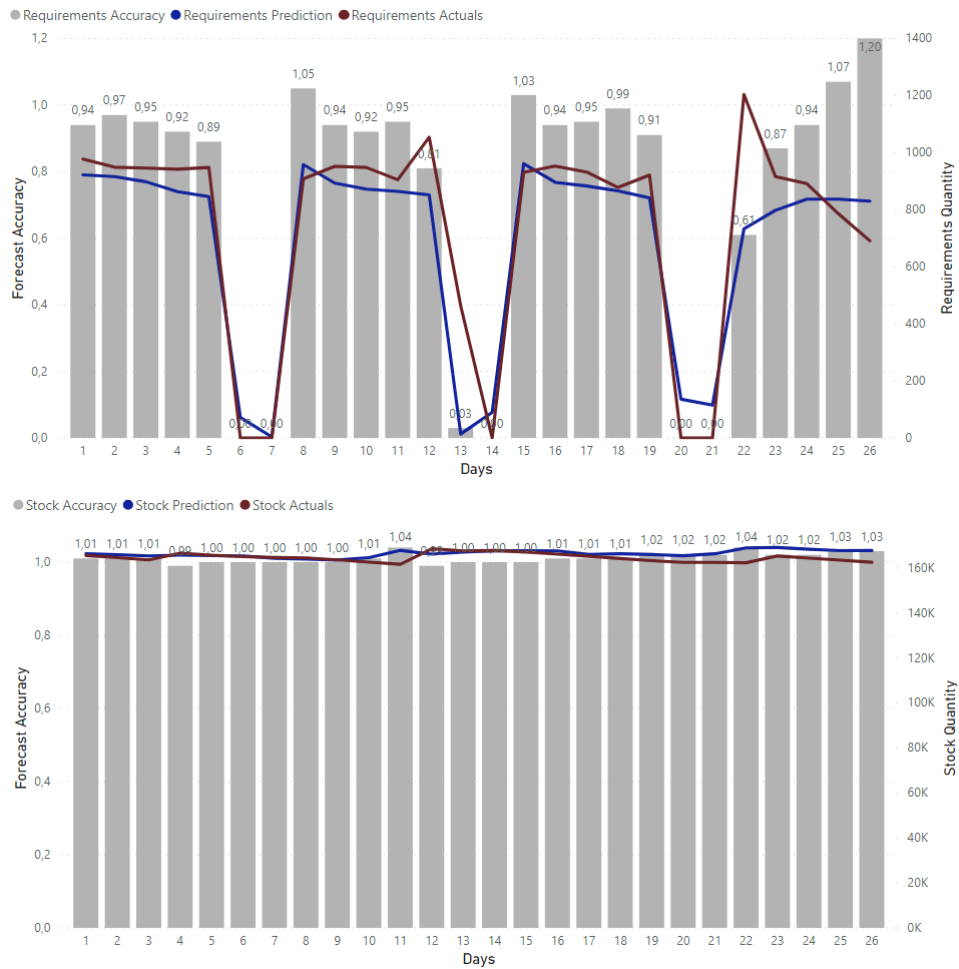


Figure 5.13: Example of results of predictions for Cluster 1 items both for stock and gross requirements.

	Acc. over Month	MAPE
Decoupled Models	3,17%	3,16%
Coupled Models	2,78%	2,88%

Table 5.9: Results with and without model mismatch for stock values predictions.

	Random Forest Regression				Multilinear Regression			
	7 days		30 days		7 days		30 days	
	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE
gross_req _qty	0,91	211,36	0,93	186,09	0,82	299,32	0,88	246,00
stock	0,98	1758,59	0,98	1612,51	0,97	2082,75	0,97	1995,39

Table 5.10: Statistics of cluster 1 predictions.

5.2.7 Cluster 2

Cluster 2 contains tail products with low volume, low value and low usage frequency. As anticipated before, the challenge here is to forecast for items that may be very marginal in the production cycle, if not completely discontinued. Also here the requirement forecast has a very low accuracy, but feeds into a much robust stock forecast. The comparison among model output accuracy would privilege a 30 days backward looking feed Multilinear regression in the requirement forecasting phase, while a 7 days backward looking feed Random Forest Regression for the stock forecasting.

	REQUIREMENTS				STOCK			
	Random Forest Regression		Multilinear Regression		Random Forest Regression		Multilinear Regression	
CLUSTER_2	7 days	30 days	7 days	30 days	7 days	30 days	7 days	30 days
MAPE	52,84%	50,35%	37,16%	35,48%	3,82%	3,92%	12,89%	10,94%
Accuracy over month	57%	53%	44%	42%	1%	2%	11%	9%

Table 5.11: Accuracy performances of cluster 2 predictions.

Indeed, choosing the decoupled models for gross requirements

and stock involves better performances. If the Random Forest Regression is applied on stock values after forecasting gross requirements values with the Multilinear Regression, there has been an improvement in terms of accuracy and MAPE (Table 5.12).

	Acc. over Month	MAPE
Decoupled Models	0,28%	3,77%
Coupled Models	1,24%	3,82%

Table 5.12: Results with and without model mismatch for stock values predictions.

Looking at statistics (Table 5.13), they have reiterated the results for stock's predictions; while, as regard the gross requirements' predictions, it seems that the decoupled models don't work well in analytical terms because of worse values associated with the application of the Multilinear Regression based on 30 previous days datapoints compared to those of the Random Forest Regression. However, the choice of the decoupled models pays much more thanks to the best guaranteed accuracy.

	Random Forest Regression				Multilinear Regression			
	7 days		30 days		7 days		30 days	
	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE
gross_req_qty	0,61	84,55	0,65	80,10	0,48	97,31	0,53	92,23
stock	0,95	699,22	0,95	678,85	0,92	924,75	0,91	951,47

Table 5.13: Statistics of cluster 2 predictions.

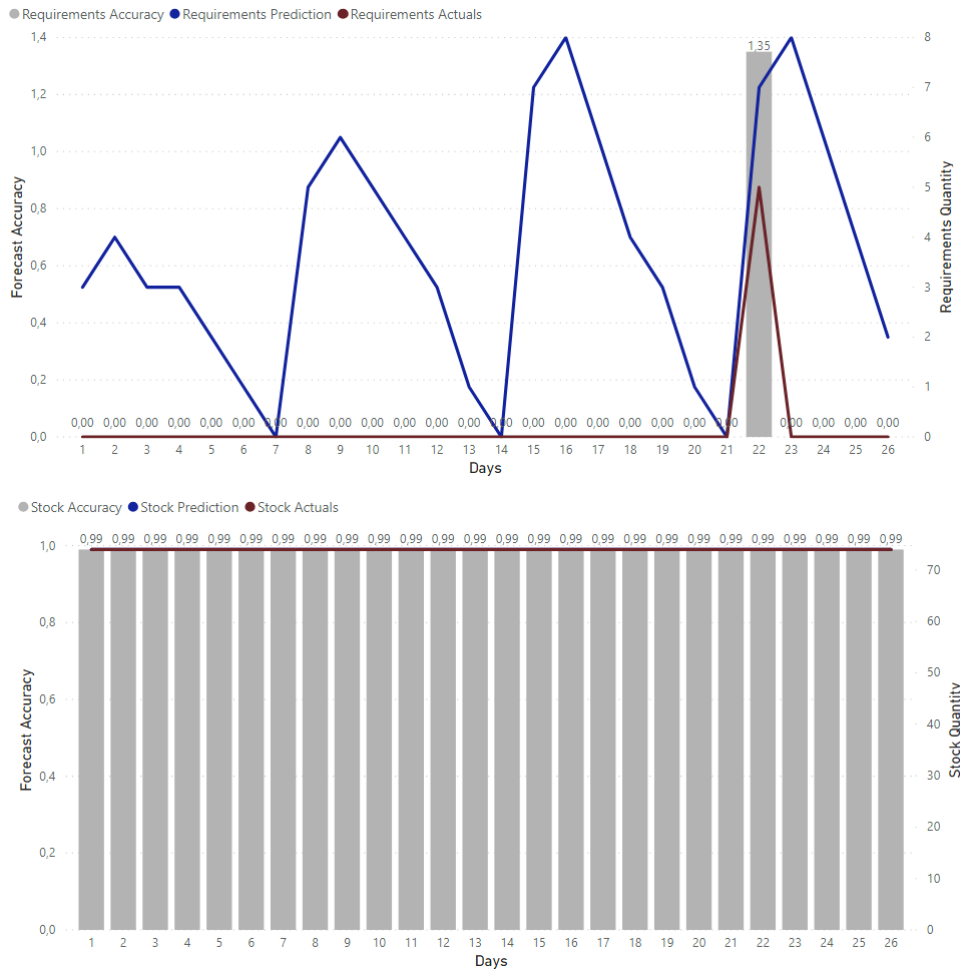


Figure 5.14: Example of results of predictions for Cluster 2 items both for stock and gross requirements.

5.2.8 Cluster 3

Cluster 3 items are similar both in terms of quantity and usage to the ones in Cluster 1. However, they are more critical as regards costs because of an higher economic contribution to company expenses. Therefore, it's important to provide the the most possible accurate forecast.

The results (Table 5.14) indicate as best choices for predictions models the Multilinear Regression for gross requirements and the Random Forest Regression for stock, both with a 30 days backward looking feed.

	REQUIREMENTS				STOCK			
	Random Forest Regression		Multilinear Regression		Random Forest Regression		Multilinear Regression	
CLUSTER.3	7 days	30 days	7 days	30 days	7 days	30 days	7 days	30 days
MAPE	15,51%	12,55%	7,77%	8,01%	7,44%	4,48%	17,23%	16,34%
Accuracy over month	11%	9%	6%	5%	3%	3%	17%	16%

Table 5.14: Accuracy performances of cluster 3 predictions.

The decoupled models are necessary because there would be an accuracy reduction if we try to align the model on the two sides. Particularly, on stock side, if we choose a Multilinear Regression the performances would be greatly affected, with accuracy loss by more than 10 percentage points. The same situation can be seen for gross requirements prediction, with a loss in accuracy if the Random Forest Regression would be selected. Moreover, Table 5.15 demonstrates that the choice of decoupled models improves precision performance in terms of accuracy computed over month and of MAPE.

	Acc. over Month	MAPE
Decoupled Models	2,26%	4,20%
Coupled Models	2,75%	4,48%

Table 5.15: Accuracy performances of cluster 3 predictions.

Returning to the results (Table 5.14), interestingly, the accuracy derived from models applications is similar, if not equal (it is the stock case), if we choose 7 or 30 days backward looking. The required accuracy, due to economic impact determined by the items of this cluster, points towards the 30 days backward looking since there is the gain of a percentage point in accuracy for gross requirements' predictions.

	Random Forest Regression				Multilinear Regression			
	7 days		30 days		7 days		30 days	
	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE
gross_req_qty	0,84	219,45	0,78	253,45	0,70	295,47	0,74	278,77
stock	<0	8696,39	0,92	1378,26	<0	5207,35	0,37	3981,27

Table 5.16: Statistics of cluster 3 predictions.

Lastly, Table 5.16 confirms that the best choice as regard the stock prediction is obtained through the Random Forest Regression based on 30 previous days datapoints. Instead, as in the case of the Cluster 2, the best model for gross requirements prediction in term of accuracy isn't that from statistics' point of view. However, as already said, this is a critical cluster concerning the economic impact determined, so it must be preferred the decoupled models, referred to above, in order to satisfy the business requirement of more precision.

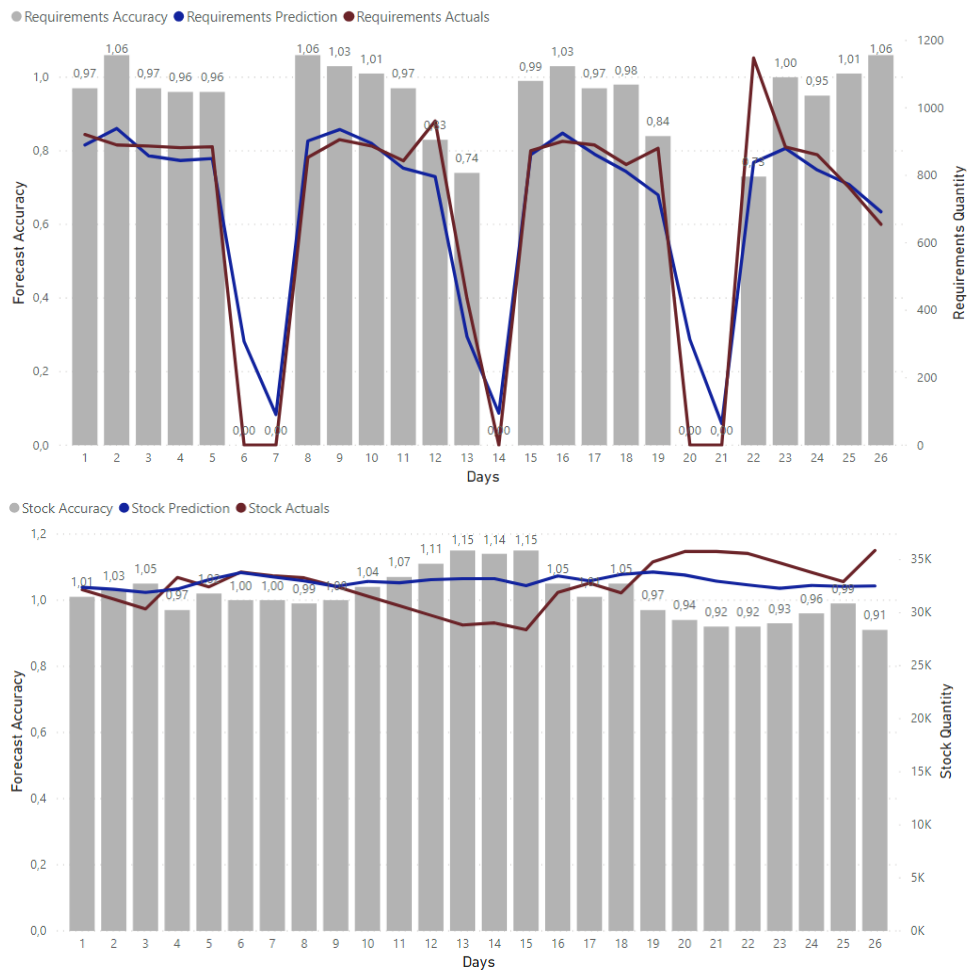


Figure 5.15: Example of results of predictions for Cluster 3 items both for stock and gross requirements.

Chapter 6

Conclusions

This study looked at a key supply chain challenge, the struggle to reach a better demand and stock forecast, with the aim of defining the best possible approach to a real manufacturing business case, by applying state of the art data mining techniques.

As a first finding, it acknowledged the opportunity of treating different products differently. The product clustering has been defined with an unsupervised learning approach, exploiting k-means algorithm. This led to 4 clusters identification:

- **Cluster 0:** high volume, low value, high frequency products
- **Cluster 1:** medium volume, medium-low value, high frequency products
- **Cluster 2:** low volume, low value, low frequency products
- **Cluster 3:** medium volume, high cost, high frequency products

The forecast phase explored two types of modeling, Multilinear Regression and Random Forest Regression, both testing a different test set feeding option (7 backward days and 30 backward days), for each cluster.

The output comparison led to the following evidence:

- For **Cluster 0** products, Random Forest Regression based on 30 previous days datapoints has to be privileged for both the Requirements and Stock forecast, as it is leading to a MAPE of 28% and 3%, respectively.
- For **Cluster 1** products, 30 days backward looking Multilinear Regression for requirements and 30 days backward looking Random Forest Regression for stock forecast has to be chosen, leading to a MAPE of 11% and 3%, respectively
- For **Cluster 2** products, 30 days backward looking Multilinear Regression for requirements and 7 days backward looking Random Forest Regression for stock forecast, leading to a MAPE of 35% and 4%, respectively
- For **Cluster 3** products, 30 days backward looking Multilinear Regression for requirements and 30 days backward looking Random Forest for stock forecast, leading to a MAPE of 8% and 4%, respectively

The overall picture shows that Multilinear Regression has worked better for the requirements forecast, with the exception of Cluster 0 and Cluster 1 products, even if in the latter case the Multilinear Regression leads to a good accuracy too. Instead, Random Forest Regression worked better for stock forecast. Also, feeding last 30 days in the model leads to better accuracy results. The only exception here was Cluster 2, where 7 backward days regression has been privileged, but 30 days algorithm results were very close. The study also confirms forecasting demand is much more difficult versus forecasting stock, due to a much higher volatility.

Despite having reached interesting results, it is important to understand what could be areas of improvements, especially when dealing with forecast algorithm reactivity.

The Random Forest Regression, as an ensemble model, by nature, is less data-sensitive (less variance) and also assures a higher

total average accuracy, but it cares less about the punctual one (represented by the daily forecast accuracy, in this study). Effectively, final predictions are the result of operations of averages on different Decision Tree models. At business level, in the analyzed case, that translates into a smoother stock curve prediction that allow to preserve continuity of service. Nevertheless, this may lead to a risk of lower turnover than what should be.

Improving the results achieved in this study is possible, starting from the conclusions reached above and looking at further forecasting strengthening techniques, employing complex ensemble models like AdaBoost or Gradient Boosting. They are both boosting ensemble methods which make predictions based on a set of different models, training them in a sequential way. Their peculiarity is that each learner model depends on the previous learner. In fact, each model learns from mistakes made by the previous one[18].

AdaBoost learns from the mistakes by using an allocation mechanism of weights for each data points and for the model. The latter is defined taking into account a weighted error rate that is the number of wrong predictions out of total. The higher is the weight of the model, the more accurately it makes predictions and the more influence it will have on the final decision. Instead, the Gradient Boosting learns from the mistakes directly referring to residual errors.

The iterative strategy used in these two boosting methods allows us to obtain a lower bias, keeping the advantage of low variance, which can be achieved with the bagging ones, too. In this way, an improved punctual accuracy can be reached, delivering even higher business value.

Bibliography

- [1] Accenture. “Big Data Analytics in Supply Chain:Hype or Here to Stay?” In: *Accenture Global Operations Megatrends* (2014).
- [2] Ashraf Mohammad Salem Alrjoub, Muhannad Akram Ahmad. “Inventory management, cost of capital and firm performance: evidence from manufacturing firms in Jordan”. In: *Investment Management and Financial Innovations* (Oct. 2017).
- [3] Batini Carlo, Scannapieco Monica. *Data Quality: Concepts, Methodologies and Techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- [4] Capkun Vedran, Hameri Ari-Pekka, Weiss Lawrence A. “On the relationship between inventory and financial performance in manufacturing companies”. In: *International Journal of Operations & Production Management* (July 2009).
- [5] Cherukuri Madhu Babu, Ghosh Tamoghna. “Control Spare Parts Inventory Obsolescence by Predictive Modelling”. In: *2016 IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data* (Dec. 2016).
- [6] Arnold J. R. Tony, Chapman Stephen N., Clive Lloyd M. *Introduction to Materials Management*. Pearson Prentice Hall, 2008.
- [7] Dallari Fabrizio, Milanato Damiano. “Pianificazione e gestione dei materiali - Dimensionamento corretto delle scorte di sicurezza”. In: *Centro di Ricerca sulla Logistica LIUC Università Cattaneo* (Jan. 2013).
- [8] Davenport Thomas H., Patil J. “Data Scientist: The Sexiest Job of the 21st Century”. In: *Harvard Business Review* (Oct. 2012).
- [9] Davenport Thomas H., Patil J. “Efficiency of K-Means Algorithm and various hierarchical clustering methods in the inventory classification”. In: *Harvard Business Review* (Oct. 2012).
- [10] De Mauro Andrea. *Big Data Analytics - Analizzare e interpretare dati con il machine learning*. Apogeo, 2019.

- [11] Deloitte. “Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications”. In: *Key findings from Deloitte’s Analytics Advantage Survey* (2013).
- [12] Deloitte. “Supply Chain Analytics. The three-minute guide”. In: (2012).
- [13] Chain of Demand. *How Predictive Analytics Will Change the Supply Chain*. 2018. URL: <https://www.chainofdemand.co/how-predictive-analytics-will-change-the-supply-chain/>.
- [14] Elsayed Khaled, Wahba Hayam. “Reexamining the relationship between inventory management and firm performance:an organizational life cycle perspective”. In: *Future Business Journal* (June 2016).
- [15] Eroglu Cuneyt, Hofer Christian. “Inventory types and firm performance: Vector autoregressive and vector error correction models”. In: *Journal of Business Logistics* (Sept. 2011).
- [16] Hopkins Michael S., LaValle Steve, Balboni Fred. “The new intelligent enterprise: 10 Insights: A First Look at The New Intelligent Enterprise Survey”. In: *MIT Sloan Manage* (Oct. 2010).
- [17] Hazen Benjamin, Boone Christopher, Ezell Jeremy, Jones-Farmer L. Allison. “A Prediction-Based Inventory Optimization Using Data Mining Models”. In: *International Journal of Production Economics* (Aug. 2014).
- [18] Chen Lujing. *Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained*. 2019 . URL: <https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49de2725>.
- [19] Claessens Maximilian. *Characteristics of the product life cycle stages and their marketing implications*. 2017. URL: <https://marketing-insider.eu/>.
- [20] Metzler Lloyd A. “The Nature and Stability of Inventory Cycles”. In: *The Review of Economics and Statistics* (Aug. 1941).
- [21] Obermaier Robert, Donhauser Andreas. “Zero inventory and firm performance: a management paradigm revisited”. In: *International Journal of Production Research, Department of Accounting and Control, University of Passau, Passau, Germany* (Aug. 2012).
- [22] O’Reilly Charles A. “Variations in Decision Makers’ Use of Information Sources: The Impact of Quality and Accessibility of Information”. In: *The Academy of Management Journal* (Dec. 1982).

- [23] Redman Thomas. “The impact of poor data quality on the typical enterprise”. In: *Communications of the ACM* (Feb. 1998).
- [24] SAP. *Documentation*. 2016. URL: https://help.sap.com/doc/saphelp_sfin100/1.10/en-US/36/57b953cccbb34ce10000000a174cb4/frameset.htm.
- [25] ShapiroJeremy F., Wagner Stephen N. “Strategic inventory optimization”. In: *Journal of Business Logistics* (Sept. 2009).
- [26] Souza Gilvan. “Supply chain analytics”. In: *Business Horizons* (Sept. 2014).
- [27] Hopp Wallece, Spearman Mark. *Factory Physics: Foundations of Manufacturing Management*. Waveland Pr Inc, 2008.
- [28] Xiaoxiao Guo, Chang Liu, Wei Xu, Hui Yuan, Mingming Wang. “A Prediction-Based Inventory Optimization Using Data Mining Models”. In: *2014 Seventh International Joint Conference on Computational Sciences and Optimization* (Sept. 2014).
- [29] Brandimarte Paolo, Zotteri Giulio. *Logistica di distribuzione*. CLUT, 2004.