LUISS

Department of Business and Management

Marketing Analytics and Metrics

Course: Statistics for Marketing

The Use of Traditional Business Metrics vs AI: Machine Learning – The Ferragamo Case

Advisor

Prof. Francesco Salate Santone

Co-Advisor

Prof. Michele Costabile

Prof. Fabio Possamai

Candidate

Stefano Landolfi 782191

TABLE OF CONTENTS

CHAPTER 1: OVERVIEW AND REFERENCE CONTEXT

1.1 OVERVIEW AND REFERENCE CONTEXT

In recent years, driven by continuous waves of innovation, the fashion industry has undergone profound transformations both organizationally and in terms of consumer expectations. E-commerce has now become an essential component for every luxury brand, facilitating more direct and efficient interactions with consumers. While this digital shift has significantly enhanced communication and simplified purchasing processes, it has also expanded and intensified market competition, giving rise to a truly global playing field. Today, every company in the sector can showcase its products to a worldwide audience, while consumers have access to an unprecedented variety of choices.

In the luxury segment, known for its ephemeral character**,** e-commerce has evolved into a fundamental strategic lever. But, Brands, having overcome their initial reluctance, are increasingly adopting an omnichannel approach that integrates online and offline experiences. This model combines the convenience of digital platforms with the exclusivity of physical retail spaces, which continue to serve as essential touchpoints for conveying the identity and values of luxury brands. Leveraging big data analytics and advanced technologies, luxury firms are now pursuing highly personalized and customer-centric marketing strategies, fully exploiting the potential of virtual environments. (Y., 2023)

Before analyzing the Fashion Industry in detail, it is important to have a clear definition of the term. Collen McDowell, from the Costume Society, describes fashion as a “form of art…that reacts more quickly to any kind of social and cultural nuance” (McDowell C. , 2000). We can therefore define fashion as a very fast-changing process, a process that is certainly appealing to customers, with all the novelties it brings, but which forces companies in the sector into constant and rapid adaptation. The result is an increasingly fast

and dynamic market, where staying ahead of competitors becomes a key factor in maintaining one's relevance within the industry**.** All the dynamics outlined above suggest that one of the most critical factors in this type of market is time-to-market. That is, the time required to develop a product and launch it on the market. It is a crucial KPI, as a rapid market entry allows companies to anticipate competitors, respond more effectively to market demand, and seize new opportunities. Reducing it means greater agility and competitiveness in a dynamic environment. Although this may seem straightforward, the luxury sector is a perfect example of how complex it actually is: before launching a product, a typical company invests months in research and analysis, since even the smallest mistake can be fatal to the planned sales performance of that product. Excess inventory, an error in product or market analysis, or worse, a negative review from customers can all result in losses, not only for the current sales cycle but also for future ones. Therefore, being able to plan everything with extreme precision and in the shortest possible time becomes a company's top priority.

Looking more closely at today's luxury market, it becomes clear that the industry is undergoing a deep crisis. Very few companies are currently able to achieve significant results and creating value around products in such a competitive and saturated environment has become increasingly difficult. This stands in stark contrast to previous decades, when the offering of high quality, unique products was far more limited, and the role of the designer held considerable influence. One of the many reasons may be that, over the past two years, the market has lost approximately 50 million consumers, partly due to price increases that have driven away middle income buyers, as reported by Vogue Business in one of its articles. (Afonso, 2008)

However, the most significant challenge the luxury market has had to face, particularly in the aftermath of the COVID-19 period is fast fashion. Fast fashion rapidly produces low-cost versions of high-end designs, making similar styles accessible to a much broader audience. This phenomenon can dilute the aura of exclusivity traditionally associated with luxury brands. (W.

H. Thejani Madhuhansi, 2025), This goes hand in hand with the growing popularity of so-called “quiet style,” a trend characterized by the absence of visible branding on clothing and a deliberate move away from overt brand display.

These represent two completely different schools of thought: on one side, the pursuit of exclusivity, luxury, and craftsmanship; on the other, the serial production model of fast fashion. These contrasting approaches to product development and business strategy have become even more evident in recent years with the rise of artificial intelligence (AI) e much more consideration is now given by companies to Big Data. Data, beyond being fundamental for individual firms, are increasingly becoming a guide toward understanding new types of consumers and, when viewed from a supply chain perspective, a clear source of value creation for raw material industries, an area that represents a historical cornerstone of the Italian market. While fast fashion has rapidly embraced advanced technologies including artificial intelligence (AI) and data analytics, to optimize operations and respond to market dynamics, many luxury brands, such as Ferragamo, have shown greater caution in implementing such innovations. This thesis aims to analyze the current state of the sector, highlighting how AI is being used, the improvements achieved, and the limitations encountered. (Shoaib, 2024) Moreover, it will examine the current situation of Ferragamo, with particular focus on its performance in recent years, investigating the potential lack of adoption of AI for predictive, descriptive, and prescriptive analytics. Finally, the thesis will propose algorithms, with particular emphasis on a predictive algorithm a topic currently under discussion within the company which could be implemented to improve overall performance. The study will also assess how traditional business metrics can coexist with new AI-based tools.

1.2 SALVATORE FERRAGAMO: COMPANY PROFILE

Founded in 1927 in Florence, Salvatore Ferragamo is one of the most iconic historic brands of Made in Italy, synonymous with craftsmanship, technical

innovation, and timeless luxury. The company was born from the vision and creative flair of its founder, Salvatore Ferragamo, famous for having crafted custom shoes for some of Hollywood's biggest stars in the 1920s and 1930s, and it has gradually established itself as a benchmark in the international luxury fashion landscape. Today, Ferragamo is a global group active in the footwear, leather goods, ready-to-wear clothing, and men's/women's accessories sectors, with a widespread presence in over 90 countries, a network of more than 400 mono-brand stores, and a strong commitment to balancing artisanal tradition and stylistic experimentation.

The creative and production core of the brand remains firmly rooted in the Florentine territory, where the headquarters and the Osmannoro management center are located, which also hosts one of the most advanced logistics hubs in Europe in the fashion industry. The Ferragamo Group is also distinguished by a brand positioning that is highly consistent with the values of authentic luxury, sustainability, and aesthetic discretion, elements that over time have shaped a solid and recognizable identity. This balance between heritage and innovation has been particularly evident in recent years, also thanks to a strategic relaunch process led by new managerial figures.

Salvatore Ferragamo is currently undergoing a critical transition phase, marked by negative economic results but also by significant strategic renewal efforts. The 2024 fiscal year, according to official Ferragamo Group sources (Ferragamo SpA, 2024), closed with a 10.5% drop in revenues, bringing consolidated turnover to €1 billion. The deterioration in performance was also reflected in profitability: gross margin fell by 11.8%, EBITDA dropped to €215 million, and adjusted EBIT plummeted from €79 million to €35 million. Net income shifted from a positive €26 million in 2023 to a net loss of €68 million in 2024. These are clear signs of difficulty, within an international context that remains fragile, dominated by geopolitical tensions, reduced consumption in mature markets, and persistent cost pressure across the supply chain. The group has aimed for a structural improvement of its distribution channels, prioritizing the strengthening of the direct-to-consumer (DTC) channel, restructuring physical retail, and rationalizing wholesale.

1.3 THE ROLE OF MERCHANDISE AND PLANNING

Within this framework lies the work carried out in the Merchandising Planning department, the business area responsible for the quantitative and qualitative planning of the product assortment, with specific attention to distribution by channel and geography. Merchandising Planning represents a strategic hub within Ferragamo's operational architecture, as it interconnects the creative choices of design with the commercial needs of the markets, two worlds that are often in contrast, defining the guidelines for aligning supply with demand, aesthetics with performance.

The work conducted during the internship, and which is the subject of this thesis, focused precisely on this operational dimension, with the goal of exploring how the integration of predictive and data-driven tools can help strengthen the quality of allocation decisions, in a context where individual intuition, although fundamental, is not always sufficient to ensure consistency, efficiency, and responsiveness at the global level. The role within Planning today represents a highly relevant function in the corporate landscape of a luxury company. The analysis being carried out is fundamental for the decisions the company must make across the seasonal collections. It starts with the study of market demand, which is essential in a constantly changing environment. However, it is important not to generalize. As previously mentioned, Ferragamo operates across more than seven regions, each characterized by different demand profiles. Every region has a unique clientele, and for this reason, highly localized analyses are often conducted. On one hand, the company tries to meet the various regional requests; on the other hand, Ferragamo is obliged to maintain a consistent brand image, that is, a global offering that reflects brand values in terms of both product and store experience.

This is precisely where one of the main issues arises that the planning departments have had to face: the allocation of SKUs across regions. During the period leading up to a new collection, decisions must be made regarding how many and which SKUs to propose, following extensive analysis, mostly based on historical data. Today, within Ferragamo, thanks to a clustering

algorithm, it is easier to determine how many SKUs to allocate, but the greater challenge lies in knowing *which ones* to choose.
Using real data provided by the company and spanning multiple consecutive seasons, the analysis focused on the handbag product category, selected for the completeness of its historical data and for the strategic importance that handbags hold in the brand's commercial and identity positioning. The complexity in managing SKU lifecycle, the geographic fragmentation of the retail network, and the high seasonality of demand made this segment the ideal candidate to test a new approach to forecasting and optimizing assortment decisions.
Through this study, an effort was made to build a bridge between the strategic dimension of the Ferragamo brand and the analytical concreteness offered by data, demonstrating that even in the world of luxury, which has historically relied more on experience and taste, there is room for a cultural and technological transformation oriented toward the conscious use of artificial intelligence in product management.

1.4 OBJECTIVES

The primary objective of this thesis is trying to develop a predictive algorithm and try to demonstrate that it will work better than a traditional operation in Ferragamo, designed to optimize the selection of SKUs (Stock Keeping Units) and to be introduced in individual retail locations, distributed across the different commercial regions in which Ferragamo S.p.A. operates currently divided into seven geographical areas. This study falls within the broader context of strategic product distribution planning, with a particular focus on

the concept of time-to-market, which will serve as a key indicator to measure the benefits of implementing the new system.

The proposed algorithm will be based on the analysis of internal databases provided directly by the company, containing historical sales data, product performance metrics, and other relevant operational variables. By studying these data, the thesis aims to identify recurring patterns and significant relationships, with the goal of developing a more efficient SKU allocation system tailored to the needs and specific characteristics of each region and store.

In parallel, a critical analysis will be conducted on the methods currently employed by Ferragamo for demand forecasting and assortment definition, with the intent of identifying potential areas for improvement.

This study will also explore the relationship between the fashion luxury sector and the adoption of AI and machine learning processes, examining how these technologies have already been partially integrated, and evaluating the specific advantages and limitations within this sector. Furthermore, comparative analyses will be conducted on similar cases from other industries where machine learning-based planning has become a cornerstone of corporate strategy.

1.5 RESEARCH GAP

The research gap addressed in this thesis concerns the near-total absence of empirical studies analyzing the application of predictive models, such as regression and classification, to optimize SKU distribution within the luxury fashion industry. While academic literature has primarily focused on fast fashion or e-commerce, contexts characterized by high product turnover, centralized stock, and real-time customer feedback, there is a clear lack of

work exploring the decision-making complexity typical of luxury maisons, where product distribution is planned months in advance, across international markets, and where the underperformance of a single SKU in a specific region can jeopardize broader brand positioning strategies.

Moreover, predictive assortment planning for physical retail stores, particularly in contexts where SKU availability is limited and guided more by brand image than by volume, is still a largely unexplored area from an algorithmic standpoint. In light of this, the present research aims to bridge this gap by proposing a hybrid model that, based on historical data provided by Ferragamo, is capable of identifying which products should be distributed to which stores, thereby improving decision-making efficiency without compromising the brand's aesthetic and strategic coherence.

This gap is therefore twofold: on one hand, it reflects the absence of machine learning applications for SKU forecasting in the luxury fashion sector, and on the other, it highlights the lack of practical approaches that leverage real corporate data in a data-driven framework within organizations traditionally oriented toward intuition and craftsmanship.

1.6 RESEARCH METHODOLGY

To achieve the objectives of this thesis, a structured and sequential research methodology was adopted, combining both theoretical investigation and practical experimentation with real-world corporate data. The starting point was an in-depth review of academic literature concerning the use of machine learning in the fashion and retail sectors, with particular attention to studies on predictive assortment planning, SKU optimization, and AI applications in supply chain and merchandising processes. This phase helped establish the theoretical foundations of the project and identify the research gap discussed in the previous section.

The empirical part of the research was conducted using a dataset provided directly by Salvatore Ferragamo S.p.A., comprising sales, stock, and product information for the Handbags category over the years 2022, 2023, and 2024, across all the regions where the company operates. A detailed data cleaning and preparation process was performed entirely in Python, using Jupyter Notebook as the development environment. This phase included the identification and treatment of missing values, standardization of categorical variables, removal of inconsistencies, and restructuring of the dataset, this phase was essential to enable comparative analysis and predictive modeling.

Following the preprocessing phase, the research focused on the development of model ….

The model training and evaluation were conducted using the historical data collected over the three-year period, and particular attention was paid to feature selection, including product characteristics (e.g., color, material, line and performance indicators from previous years. Additional variables were engineered to enrich the models.

The final step of the methodology involved the analysis of the results from both a technical and business perspective, to assess the applicability of the proposed solution in a real corporate environment. Special focus was placed on the potential benefits of the model in terms of improved decision-making, reduction of overstock or understock risk, and acceleration of time-to-market.

1.7 THESIS STRUCTURE

This thesis is structured into five chapters, each of which contributes to the development of a predictive model aimed at improving SKU distribution in the

luxury fashion sector, with a focus on Ferragamo's handbag category. The structure is as follows:

Chapter 1 – Overview and Reference Context

The first chapter introduces the topic, providing a comprehensive overview of the research framework and the fashion luxury market. It outlines Ferragamo's corporate profile, the strategic role of merchandise and planning, the research objectives, the identified gaps in literature, and the methodological approach adopted for this thesis.

Chapter 2 – AI in Fashion Luxury and Why It Could Prove Important

This chapter explores the growing importance of artificial intelligence in fashion, analyzing the industry's current challenges, the complexity of SKU allocation decisions, and the opportunities of adopting predictive models. It includes a market overview, consumer behavior studies, and discusses how data-driven decision-making can impact Ferragamo's operations.

Chapter 3 – Data Description

The third chapter provides a detailed description of the dataset used, the key variables selected, and the preliminary cleaning steps. It focuses on the structure of the data, the transformation process implemented through Python, and the challenges encountered during the preprocessing phase.

Chapter 4 – Predictive Model

This core chapter presents the exploratory analysis performed after data cleaning and introduces two predictive approaches: binary classification and Poisson regression. It explains the modeling rationale, the logic behind the chosen techniques, and includes a simulation on fictitious SKUs to validate the coherence of the predictive system.

Chapter 5 – Final Application on Real Ferragamo Data: Poisson Prediction and Binary Classification of Handbag SKUs

The final chapter applies the developed models to real Ferragamo handbag data, interpreting the strategic significance of the results. It evaluates performance across different regions, identifies SKUs suitable for Global Core Assortment (GCA), and discusses how the model can support decision-making to reduce time-to-market. The chapter concludes with a reflection on possible limitations and future developments.

Conclusions

The thesis ends with a general reflection on the strategic potential of AI in fashion planning, reinforcing the need to move towards a more structured, data-driven approach in the luxury industry.

CHAPTER 2 – AI IN FASHION LUXURY AND WHY IT COULD PROVE IMPORTANT

2.1 INTRODUCTION TO ARTIFICIAL INTELLIGENCE IN THE CONTEXT OF FASHION LUXURY

In recent years, artificial intelligence (AI) has progressively established itself as one of the most promising and cross-functional technologies in the field of industrial innovation, finding applications even in sectors traditionally considered distant from technological contexts, such as fashion. Within the luxury segment, however, AI has so far been adopted by a limited number of companies, primarily for marketing purposes as seen in the case of LVMH's "AI Factory" to enhance the customer experience. AI is capable of simulating human cognitive abilities, such as learning from experience, recognizing patterns, adapting to new information, and making predictive inferences. Its application in the fashion industry thus opens new opportunities for efficiency, personalization, and responsiveness to market dynamics.

The integration of AI into the fashion industry extends across the entire value chain: from the creative and design phases to production, logistics, marketing, and customer experience. The most relevant technologies in this context include machine learning, deep learning, computer vision, and natural language processing (NLP).

In terms of operational functions, the fashion luxury sector has yet to make significant progress. For instance, among the most impactful components of artificial intelligence, machine learning stands out as a key tool that enables computer systems to learn from large volumes of historical data to forecast future behaviors a crucial capability for demand prediction and assortment planning. Nevertheless, few leading companies in the fashion industry currently treat it as a core strategic focus. "Marketers and researchers are far from having a thorough understanding of the broad range of opportunities ML

applications offer in creating and maintaining a competitive business advantage" (De Mauro, 2022) As noted by A. De Mauro in one of his studies on the potential of machine learning within marketing and beyond, this technology holds considerable promise. Another relevant example is deep learning, a subfield of machine learning, which is primarily employed by companies for the analysis of complex images. This allows organizations to identify emerging trends an AI-driven analytical approach that could prove crucial in a highly dynamic market, as suggested by recent studies published in the *International Journal of Management & Entrepreneurship Research* (Ogundipe, 2024).

Fashion, characterized by rapid cycles and high demand volatility, could view artificial intelligence as a strategic tool to support innovation, optimize costs, and enhance competitiveness. However, it is essential to address the challenges associated with AI implementation. The luxury fashion industry today represents one of the most dynamic and strategic sectors of the global economy. According to the latest Bain & Company report (McDowell C. (., 2024)the personal luxury market has reached a value of over 362 billion euros, driven by growing sales in Asia and North America. However, following the post-pandemic euphoria, growth prospects in mature markets such as Europe and the United States have recently cooled, with estimates indicating a slowdown in the range of +3% to +5% CAGR through 2026, in the face of an uncertain geopolitical context and increasing consumer attention to sustainability, personalization, and omnichannel shopping experiences (Bain & Company, 2024). In this evolving scenario, the adoption of advanced technologies such as artificial intelligence (AI), computer vision, predictive analytics, and the Internet of Things (IoT) is profoundly transforming the way luxury brands manage their supply chains, operations, and customer relationships. Unlike fast fashion, where stock rotation logic and real-time feedback have long been optimized through intelligent algorithms, in the luxury sector technological innovation is still in a consolidation phase. However, the very structural complexity of these companies – founded on

craftsmanship, exclusivity, and global physical presence – makes the integration of data-driven systems more strategic.

According to the "AI in Fashion" report by Statista (Statista, AI in Fashion - Global Market Forecast 2022–2027., 2024), global spending on artificial intelligence solutions in the fashion sector will exceed $4.4 billion by 2027, with an estimated CAGR of 34.1% between 2022 and 2027. What stands out is that an increasing share of these investments is specifically targeting the luxury segment, where AI is seen not only as a lever for efficiency but also as a tool to preserve and strengthen the brand's identity value. For example, the LVMH group announced in 2023 the creation of an internal hub – the AI & Data Lab, with the goal of integrating predictive algorithms into merchandising, forecasting, and demand management processes. Similarly, Kering has developed a proprietary infrastructure to optimize collection planning based on historical product performance, leveraging neural networks and nonlinear regression to segment clients and forecast trends.

Ferragamo too, as highlighted in the case study of this thesis, has initiated a data-driven transition within the Merchandising Planning department, which will be explored in the next chapter, aiming to overcome its traditional reliance on purely intuitive decisions. In a context where local demand variability and the internationalization of retail channels make it increasingly difficult to anticipate the success of a specific SKU, promising results have emerged from a clustering process designed to determine how many SKUs to send to each store.

Equally important is the role of the Internet of Things (IoT), now increasingly integrated into the logistics processes of luxury fashion. According to McKinsey's "The State of Fashion Technology" report (Company., 2023), over 51% of European luxury brands have already introduced RFID sensors, NFC technologies, or digital tags in their products, not only to improve inventory monitoring and supply chain traceability but also to offer personalized post-sale experiences. The case of Burberry, which developed the "Voyage" project in collaboration with IBM, involved a system to track the entire life cycle of a

garment using cloud and blockchain technologies, from production to distribution. Similarly, Prada, Moncler, and Ferragamo, the latter known for its warehouse located at its headquarters in Osmannoro (FI), are experimenting with smart warehouses that use IoT flows to optimize seasonal rotation, reduce unsold inventory, and generate real-time insights on regional performance.

On the marketing and customer relationship management side, AI is also increasingly present in the luxury sector. Unlike the mass-market approach typical of fast fashion, here artificial intelligence is used to create hyper-personalized in-store or app-based experiences, analyzing the purchasing behaviors of top-spending customers, suggesting outfit combinations based on aesthetic preferences detected through computer vision, or even predicting customer lifetime value for loyalty purposes. Recent research demonstrates the effectiveness of predictive models in the luxury sector. In particular, a study by Kumar et al. (Kumar, 2024)implemented various machine learning algorithms, including decision trees, ensemble methods, and neural networks, to forecast consumer buying behavior within the luxury fashion sector. The results show a consistent increase in prediction accuracy compared to traditional methods, suggesting direct impacts on customer retention, inventory optimization, and product development strategies.

Nevertheless, the academic literature continues to highlight a significant gap in empirical studies on the development of predictive models applied to specific SKU distribution challenges in luxury brands. Unlike the fast fashion and e-commerce sectors, where AI adoption is now well established, in luxury the integration of predictive techniques with corporate decision-making processes remains marginal and poorly systematized. This study aims to address this gap through the analysis of a concrete case, Ferragamo, and the development of a model designed to support allocation decisions in a complex, qualitative, and sensitive environment such as that of luxury fashion.

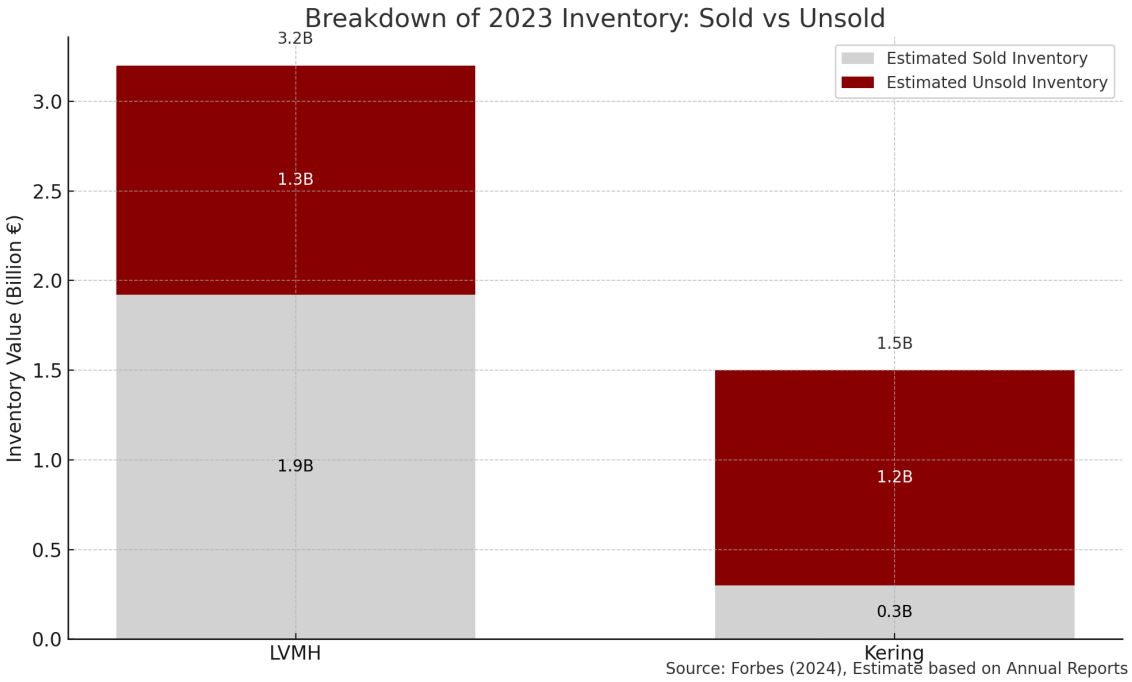
In contrast, other industries have now largely overcome the barriers to AI adoption. The following sections will explore comparable case studies drawn from market contexts similar to that of fashion luxury, reconstructing and simulating the application of various machine learning algorithms that could later inform the analysis of Ferragamo's operational processes.

2.2 PROBLEM ANALYSIS IN FASHION LUXURY

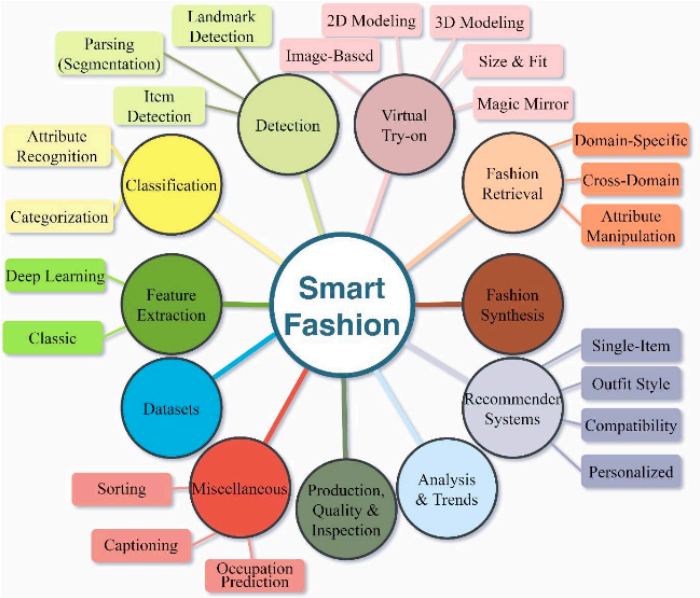
In the landscape of fashion luxury, the decision-making phase that extends from artistic creation to the actual commercialization of the product represents a delicate balance between creative instinct and rigorous operational planning. Unlike fast fashion systems, in the luxury sector the decision cycle prioritizes aesthetic content, narrative identity and exclusive experience, but it must also contend with crucial logistical, temporal and strategic challenges.

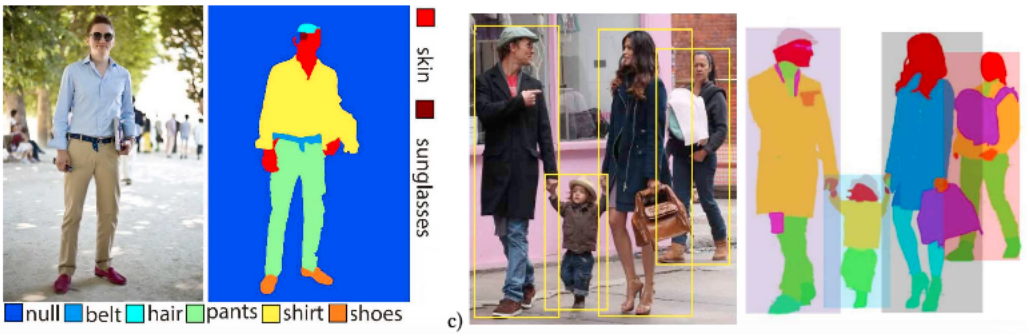
One of the main challenges concerns the long lead times, according to research by the Boston Consulting Group, fashion companies plan styles, colors and sizes with a margin of 37 to 45 weeks, or nearly ten months in advance, which corresponds to the time span between design and production. This extended timeframe makes planning highly vulnerable to rapid changes in consumer taste, emerging trends and unforeseen events, the ability to predict a product's success well in advance thus becomes fundamental. In addition to the temporal aspect, the issue of managing unsold inventory arises with urgency. McKinsey & Co. estimates that in the U.S. market alone, unsold goods reached a value of 740 billion dollars in 2023. In the luxury sector, this is compounded by the risk of brand devaluation through discounting, unlike the continuous renewal logic typical of fast fashion. For this reason, the assortment must be carefully calibrated, avoiding stock accumulation while preserving the brand's image of exclusivity.

Assortment optimization, namely the decision of which products, in what quantities and where to distribute them, also depends on geographic context and market segmentation. Luxury brands must adapt their offerings to local preferences, specific seasonalities and the identity of each individual store.



Starting from the assumption that in every global fashion company, nothing is left to chance, even if it may seem elitist and superficial, the study behind every single high fashion item is thoroughly planned. As clearly explained by Mohammadi & Kalhor (Kalhor, 2022) in an article where all the new decision-making processes on product development in fashion companies are described in detail.





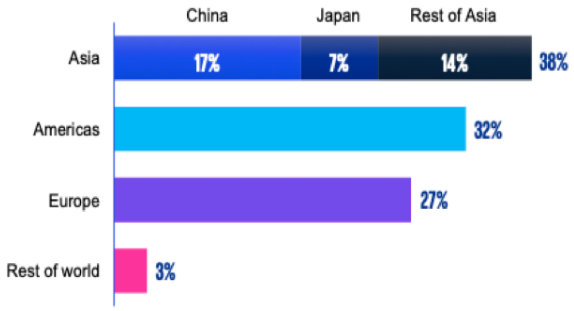
The analyses conducted both from a figurative perspective and from a market standpoint regarding the demands of new consumers are an essential element in a high fashion company. Generating value through actions that are always aligned with the fast-paced demands of consumers is the most important challenge. To give more specific examples, also related to the topics covered in this thesis, the product information available before market launch is primarily physical: product characteristics such as color, size, shape, and material, and the only numerical feature, price. However, the latter is a variable dependent on the product characteristics. All these features are decided pre-launch during months of close collaboration between the style and finance departments, with the bridge between these two figures being precisely the merchandising function.

2.3 MARKET ANALYSIS AND SPECIFIC STUDIES ON THE CONSUMER

Continuing the discussion from the beginning of the chapter, it is the market that dictates what the product will be, a market segmented according to the target audience to which the company wants to appeal. Chen-Yu, J. H., & Yang, J. H. (Chen-Yu, 2020), in an article, explore how individual characteristics, as one might expect, influence purchasing behavior. While this is relatively intuitive, the analysis becomes more complex when the reference market changes. A single consumer becomes part of a clustering process, thus joining a group of reference consumers, within a specific geographic area and seasonal timeframe. Knowing these characteristics, the product must be designed to satisfy the target consumers in seven different regions of the world, in the case of Ferragamo. Every color must be appropriate to the clothing culture of the seven regions, and so must the shape. Above all, considering all these factors, the pricing offices must define the right materials in order not to exceed a certain cost threshold and avoid increasing the price to the extent that it surpasses market expectations in certain regions.

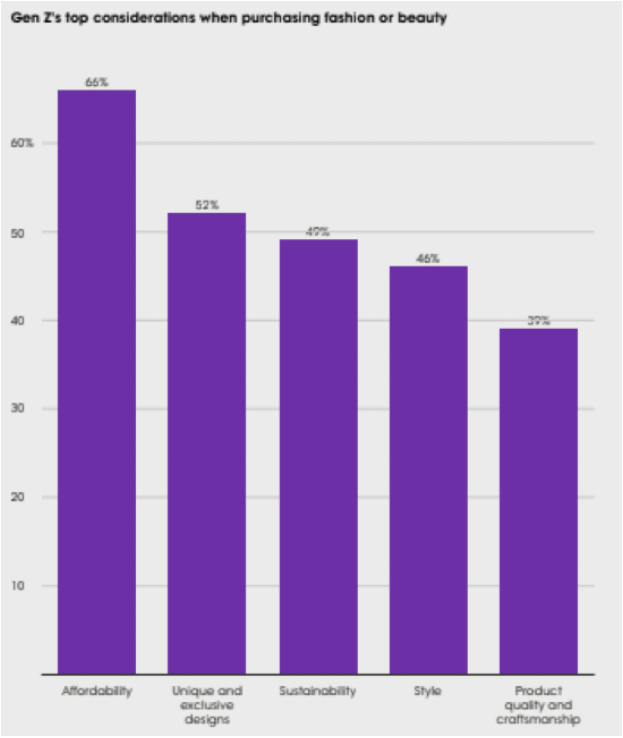
Taking more specific examples related to the thesis topic, handbags, in the case of Ferragamo, as with many other luxury brands, the shape of the bag, the color palette, and even the type of closure are decisions that are not driven purely by creative logic but derive from careful observation of customer preferences and purchasing habits in each region. For instance, in Japan and the broader Asian market, there is a marked preference for more compact formats and neutral tones, consistent with a culture of understated elegance and practical needs related to urban mobility and public transport. In North America, on the other hand, customers show more openness toward structured models, medium or large sizes, and bolder colors such as burgundy or forest green, which are well-suited to seasons like Fall-Winter.

Another emblematic case concerns the use of white: while in Western markets white is often associated with purity and minimalist elegance, in some areas of South Asia it may carry funerary cultural connotations, making it less suitable for luxury accessories intended for celebratory or social use.



For this reason, the chromatic strategy must necessarily vary from one region to another, even within the same collection.

According to the KPMG report (KPMG, 2024), the global personal luxury goods market by region is distributed as follows: Asia (38 %), Americas (32 %), Europe (27 %), and the rest of the world (3 %). These percentages highlight how the Asian market holds an increasingly significant weight within the overall sector. Following various market analyses, brands aim to understand and establish their specific positioning within each market, because, although it may seem surprising, Ferragamo's economic strength comes primarily from the European market. This supports the thesis that each market has very different demands and therefore requires distinct strategic actions.



This chart, taken from *Vogue Business 2025*, illustrates how purchasing behaviors have changed significantly. A company like Ferragamo, renowned for the quality of Italian craftsmanship, is now forced to engage with Generation Z, for whom “quality” ranks lowest in importance. With the arrival of Maximilian Davis, there has certainly been a significant step forward in the stylistic direction Ferragamo is pursuing, but this is highly constrained in order not to distort the brand’s own identity.

Therefore, Ferragamo’s main customer base remains predominantly older, placing much greater value on the origin and craftsmanship of a garment. Some academic theses identify specific consumer groups such as *Timeless Chic* (average age >35) and *Mindful Minimalists* (average age >40), characterized by strong brand loyalty, a concept far more volatile among younger generations.

2.4 DATA DRIVEN PROGRESS

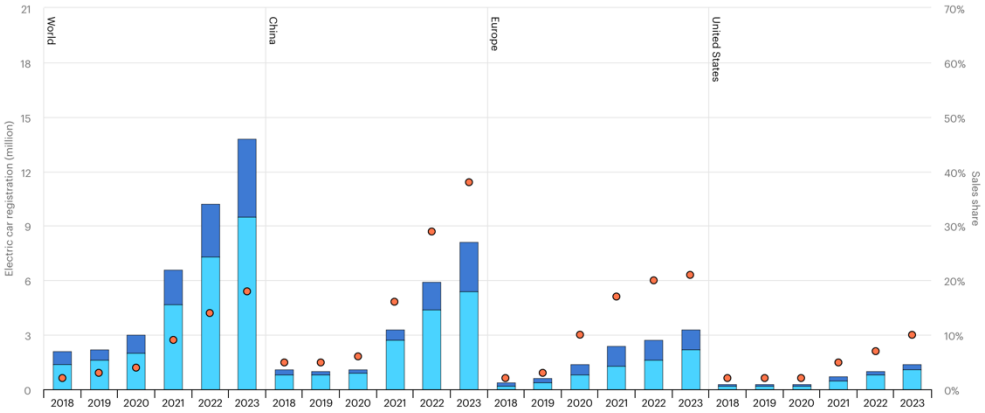
As previously explained, what has undoubtedly increased in companies in recent years is the importance placed on data analysis concerning markets, products, and competitors. Moving beyond and without focusing exclusively on the fashion luxury sector, we can find examples that are quite relevant to the study developed in this thesis. In fact, two fairly similar sectors are the automotive and beauty industries, where the creation of willingness to buy is essential. One example can be seen in the case study that led Volkswagen to cancel the launch of one of its vehicles in North America. A significant example that demonstrates the strategic value of predictive data analysis in product management is the case of the Volkswagen ID.7, an electric sedan initially intended for the North American market. Announced in 2023 and developed to compete with the Tesla Model 3, the ID.7 was a key element in Volkswagen’s electrification strategy. However, during 2024, the German

group first postponed and then cancelled the launch in the United States and Canada, reallocating resources to more responsive markets such as Europe and China. This proves the importance of a highly specific market-based analysis, rather than a generalized approach. This decision was not made arbitrarily but was instead the result of a detailed analysis of market conditions, likely supported by demand forecasting tools. Some of the key factors that influenced the decision included: the weak penetration of electric sedans in the U.S. market, where only 25% of EV sales in 2023 involved this type of model, compared to a clear dominance of SUVs and pickups; the model's ineligibility for U.S. incentives (due to assembly in Germany); and Tesla's strong position, which in 2024 accounted for over 40% of the U.S. EV market.

To assess the launch scenario, a multiple linear regression model was simulated with six quantitative variables, including factors such as price, incentives, SUV preference, competitors' market share, and macroeconomic conditions. The result was clear: the forecasted potential demand was insufficient to cover industrial, logistical, commercial, and brand positioning costs, estimated at between €49,000 and €53,000 per unit sold. This confirmed that the electric sedan segment in North America is highly unstable and unprofitable. (Stumpf, 2024)

Rather than proceeding with a risky launch, Volkswagen chose to withdraw in advance, avoiding economic losses and protecting the brand's reputation in the electric vehicle sector. This example is relevant to the present thesis, as it illustrates how the collection, modeling, and accurate interpretation of data can guide operational and strategic decisions, with a direct impact on profitability and brand strategy coherence. In the case of Ferragamo, a similar approach can be adopted to decide which SKUs to distribute, in which markets, and with what timing, significantly improving assortment management and reducing the risk of unsold inventory.

Electric car registrations and sales share in China, United States and Europe, 2018-2023



Volkswagen ID.7 Canceled For North America After All

VW has officially scrapped its plans for the ID.7 in the U.S. and Canada.

(Agency, 2024)

In this specific case, predictive algorithms such as regression were used to calculate how demand for this specific product and its characteristics would behave in response to changes in imposed incentives, a case very similar to the present study of this thesis.

As also in the case of P&G, in the cosmetics and personal care sector, Artificial Intelligence (AI) is increasingly used not only for the personalization of the customer experience but above all to optimize supply chain and logistics operations. Global investments in beauty tech technologies are expected to reach $8.96 billion by 2025, with a projected growth to $27 billion by 2034, corresponding to a CAGR above 20% (Statista, 2024)Large companies such as Procter & Gamble (P&G) have embraced AI on a widespread scale, employing

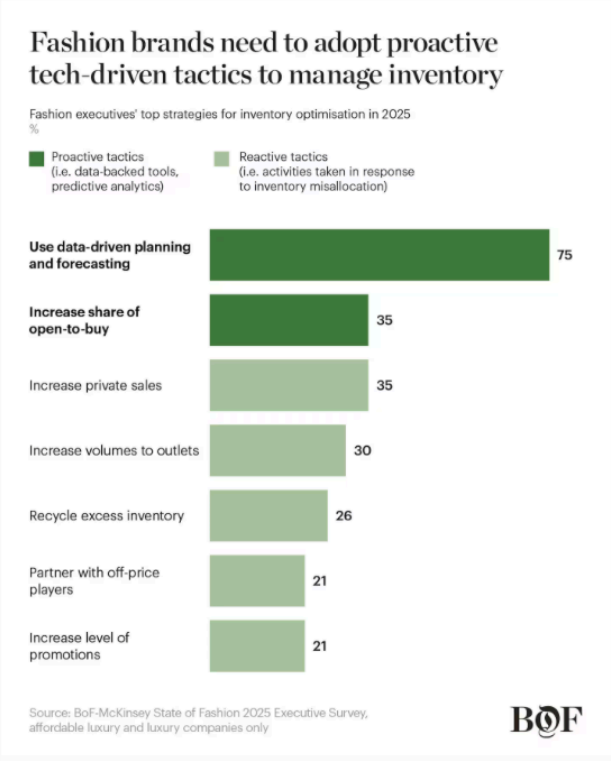
it to improve demand forecasting, reduce lead times, and strengthen collaboration with suppliers.

On the customer side, tools such as Olay Skin Advisor and Gillette Shave Advisor demonstrate how personalization through computer vision and recommendation algorithms significantly increases engagement and conversion rates. However, it is in the operational area that AI has shown the greatest strategic potential. Managing over 5,000 SKUs and more than 22,000 components in the Personal Care division alone, P&G has adopted an advanced predictive analytics system, developed in collaboration with phData and the KNIME platform, which integrates data from production, marketing, quality control, and logistics (KNIME, 2025)This system makes it possible to simulate delay scenarios, improve procurement decisions, and optimize inventory levels in real time. The result has been an estimated annual savings of between $200 and $300 million and a significant reduction in planning disruptions.

A comparable example is represented by the GE Gas Power case, in which Random Forest models were used to predict product availability dates based on variables such as contractual delivery time, approval times, and supplier reliability (Camur, 2024). In particular, the Contract Delivery Time proved to be the most relevant variable in predicting lead time, explaining and suggesting strategic levers to act upon in procurement and planning processes. The methodology employed, variable importance, one-hot encoding, training/test split, proves applicable to other sectors as well, including fashion luxury. This example, though tied to the beauty world, is particularly useful for Ferragamo, where the management of a fragmented, seasonal, and international supply chain makes product availability forecasting crucial. Applying such predictive models would allow the company to reduce time-to-market, limit overstock, and optimize resource allocation.

	Feature	Importance
41	Contract_Delivery_Time	0.572848
43	Latest_Promised_Date	0.281191
39	Product_Cost	0.036986
42	Latest_Need_by_Date	0.028271
40	Product_Quantity	0.023585
44	Approval_Time	0.019258
31	Supplier_Location_Germany	0.006352
34	Supplier_Location_USA	0.003898
37	Product_Details_OralCare	0.003343
28	Supplier_Code_S09	0.003255

Looking at these two examples, one purely statistical and the other involving the use of AI, it becomes clear how much of a difference a good data strategy makes today in any business context.



Certainly, the world of fashion luxury has made significant progress in investing in data-driven approaches, as also shown by the Bof-McKinsey (Bof-McKinsey , 2025)chart on data usage to manage inventory operations.

As mentioned earlier, the world of fashion luxury is still a world quite reluctant to rely entirely on data; some decisions are still dictated by human intellect, which does not necessarily represent a flaw. The machine only suggests what

it knows, unlike humans who most of the time invent. However, guidance could prove crucial in some circumstances.

2.5 THE NEED BEING ADDRESSED: DECISION-MAKING COMPLEXITY AND PREDICTIVE OPPORTUNITY IN SKU DISTRIBUTION

In the fashion luxury sector, every seasonal collection is the result of an extremely complex process that combines creativity, marketing, production, logistics, and sales. However, while the creative act is deeply centralized (a single design office in Florence), its commercial implementation is extremely fragmented: Ferragamo, for example, interfaces with over 400 retail stores distributed across seven macro-geographical regions, each with cultural, economic, seasonal, and structural specificities.

Within this architecture, SKU allocation, that is, the choice of which items to send to which store and in what quantity, represents one of the most critical challenges and, paradoxically, is still largely guided by intuition or individual experience rather than by structured models.

Why does this disparity (centralized creation vs. fragmented distribution) require a predictive solution? Because although creation is aligned with the brand's DNA, distribution must be modulated according to the expectations of the local consumer. Harmony is needed between the universal and the particular. Studies such as Ghauri & Cateora (2014) discuss how global companies calibrate products for local markets, not simply by adapting them but by anticipating the needs of the specific consumer.

The operational problem this thesis intends to address concerns the lack of a predictive tool that automatically supports the selection of SKUs through quantitative metrics such as units sold. Certainly, a case like this, tackled for a company of this scale, will raise many questions, since the analysis will only deal with the historical data of handbags made available by the company. A

broader view, also considering the overall market, is definitely a path to pursue, a topic that will be better addressed in the following chapters.

Returning to a scenario that the company could already undertake; we will now analyze the main differences that may arise between the two approaches.

Today Ferragamo manages SKU allocation through 3 main phases:

•**Clustering of stores**: the boutiques are grouped into homogeneous clusters based on characteristics such as sales, store size, and positioning.

•**SKU quantification per cluster**: a decision is made on how many SKUs to send to each cluster through an already active algorithm (developed using Microsoft Excel).

•**Manual selection of SKUs to be sent**: this phase remains entirely under the responsibility of planners and buyers. It is the final and crucial phase, requiring many meetings and analyses. Qualitative and quantitative metrics are analyzed and compared in a short time, and SKUs are then redefined after several meetings.

These three phases take place in the period preceding the reference collection and, in 99% of cases, buyers, as is appropriate, try to decide what is right to buy for their own region, and this is sometimes in complete disagreement with a KPI used in the company called **Global Core Assortment (GCA)**, an indicator that checks the percentage of SKUs present in all 7 regions.

This could be resolved by implementing simple guidelines, but in every company, those who invest and are operational partners will always seek to improve their own region.

Here lies the problem of why a predictive model could resolve these corporate discrepancies. A planner, beyond having some intuition from browsing

company data, has little to rely on when discussing with buyers to maintain a GCA consistent with the company. Yes, the historical data is there—once a model is created for a bag, for example, it is possible to make some very generous estimates, but nothing more. In light of this, the possible risks can be listed as follows:

•**Information overload**: each planner must manage hundreds of SKUs, dozens of stores, and dozens of variables (e.g., color, line, price, sell-through, promotional moments, stock received...).

•**Redundancy and slowness**: the analysis is performed repeatedly, with high organizational costs.

•**Individual bias**: junior or new planners may make evaluation errors, impacting distribution.

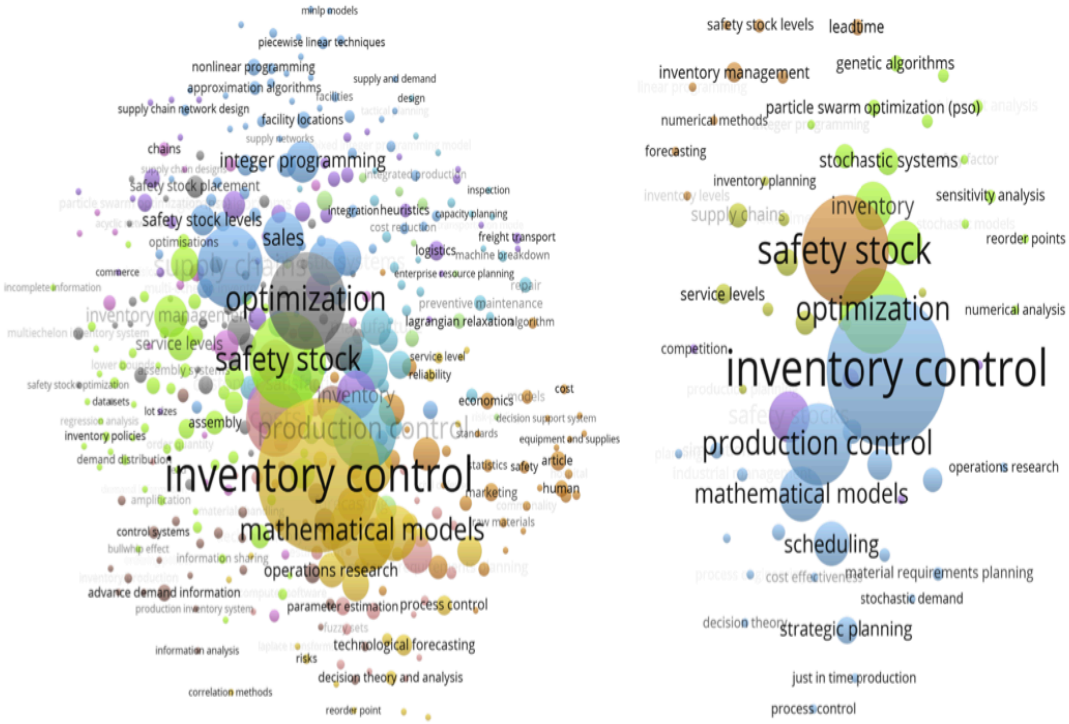
•**Strategic misalignment**: some SKUs may perform poorly in certain markets due to cultural or climatic asymmetries (e.g., colors or materials unsuitable for the Asian market or the season).

•**Unsold stock risk**: poor allocation leads to saturated warehouses and below-expectation sales.

And what if, after the clustering process, an algorithm could intervene that, even before the creation of SKUs, manages to quantify the performance that a given SKU will have in each region?

Of course, this is still just an idea, but it could create a much more streamlined product landscape, where the regions, which in the fashion world have the right to make intuitive choices, would be facilitated in their selection, allowing the company to maintain a GCA with clear guidelines and significantly reduce lead time, thus achieving optimal time-to-market. Numerous studies confirm that a data-oriented approach, specifically **machine learning**, allows for greater ease and accuracy compared to traditional methods. Spiliotis et al.

(al., 2020) – *Comparison of statistical and machine learning methods for daily SKU demand forecasting* compares the forecasting performance of ML methods (neural networks, tree-based models) with traditional statistical techniques (e.g., Croston method) on real daily demand data. ML models, especially those with cross-learning between SKUs, demonstrate higher accuracy and lower bias than traditional techniques. Haque et al. (2023) also evaluates traditional and ML models including macroeconomic variables, and in this case to a "significant reduction in lead time" is reported.



It is therefore possible to hypothesize what the advantages would be in a scenario where this algorithm is integrated into Ferragamo's decision-making process.

•Automated SKU selection: the model would propose SKUs to include based on historical sell-through, seasonality, and product variables.

•Reduction of human error: it standardizes the process and allows for conscious interventions, based on objective outputs.

•Reactivity to market signals: the model can detect anomalies (e.g., increasing unsold inventory) and suggest reallocations.

• Operational and economic efficiency: it reduces decision-making times, lead time, idle capital, and optimizes stock turnover, increasing the ROI of the collection.

It is also true, however, that beyond comparing all this with many related studies, one must consider that the data and the company addressed in this thesis are real. As widely explained previously, Ferragamo closed 2024 in the red. All new projects are on hold until the company's results improve, and a study like this would require a large number of personnel that would, in some way, change the way data is collected and processed, radically altering the entire process that the company has consolidated over the years. Investments now are mostly focused on style, as confirmed by the new AI25 releases, with the study of new materials. But here it is worth pausing briefly, because one thing should not exclude the other. Style research should not be limited by this algorithm, but rather guided, an ideal that is still poorly regarded in fashion luxury companies.

2.5.1 POSSIBLE IMPACTS ON FERRAGAMO'S BUSINESS

. Reduction of unsold inventory risk and warehouse saturation

One of the most evident benefits of introducing predictive models in the SKU allocation phase concerns the containment of unsold inventory. As shown in the chart previously developed, even leading fashion luxury groups such as

Kering and LVMH hold billions of euros in inventory each year, representing percentages between 4% and 8% of total annual revenue. In a context where unsold stock cannot be liquidated easily, due to the product's premium positioning, even a 10% reduction in stock represents a potential seven-figure saving. Ferragamo, operating on an international scale with a multi-regional system and a fragmented distribution network, is particularly exposed to this inefficiency: the misalignment between product and local demand often leads to excess stock in certain areas and stock-outs in others, increasing pressure on warehouses and sales.

. Improvement of global consistency and brand positioning

The presence of a structured model allows for greater consistency across markets, avoiding situations where the same product performs strongly in one region and poorly in another due to a subjective choice or poor data interpretation. This directly impacts brand image, customer experience, and positioning strategy. A system that takes into account product life cycle, local preferences, and marketing strategy (e.g., planned promotions, capsule collections, in-store events) ensures that brand identity is expressed homogeneously, reducing distortions.

. Optimization of SKU and regional profitability

Aligning product with demand maximizes the likelihood of full-price sales, a crucial element in fashion luxury where full-price sell-through is a fundamental KPI. A predictive algorithm can identify regional demand patterns for price and product type, helping reduce the use of markdowns and late promotions. According to a study published by Ransbotham et al. (SAM RANSBOTHAM, 2021), companies implementing predictive tools in pricing and inventory management processes report an average increase in operating margin of 5–8%.

. Faster time-to-market and greater seasonal adaptability

Another benefit concerns the possibility of shortening the decision-making cycle: if today it takes weeks of analysis and discussion to decide which SKUs to send, a data-driven model allows for testing multiple scenarios in real time, enabling faster time-to-market and a greater ability to adapt to external shocks (e.g., weather, geopolitics, local events). The effect is not just operational: according to McKinsey (2023), fashion companies that reduce time-to-market by 3–4 weeks increase seasonal sell-through by 10–15%, reducing stock accumulation and improving availability of core sizes during peak season moments.

. Simulations and what-if scenarios: a strategic support to governance

Finally, a well-designed predictive system can become not only an operational tool but also a simulation platform to support strategic decisions. Consider, for example:

•evaluating the impact of a reduction in buying budgets,

•testing new geographical clustering logics,

•simulating an omnichannel strategy that includes new formats (e.g., outlet, travel retail),

•monitoring the response to localized promotions.

CHAPTER 3: DATA DESCRIPTION

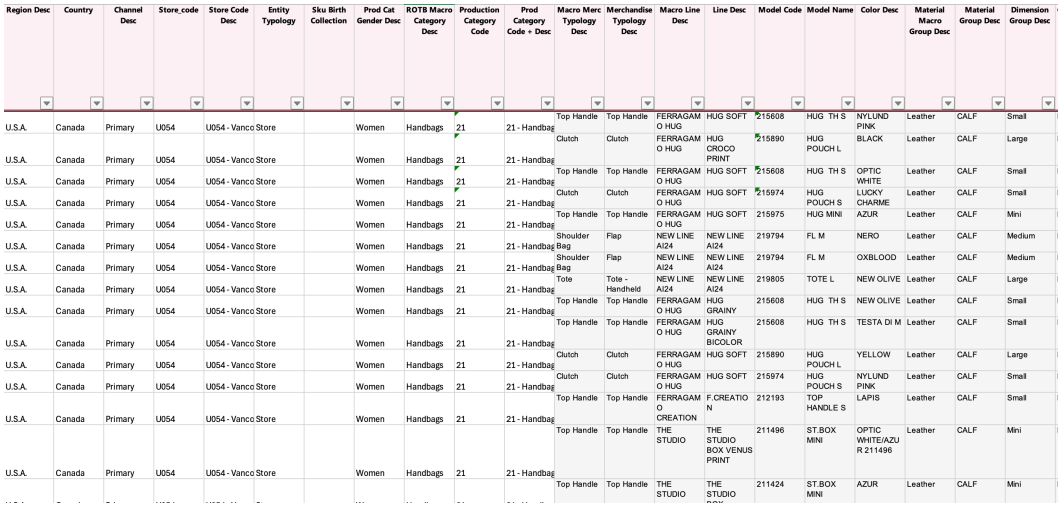
3.1 DATASET DESCRIPTION: STRUCTURE, VARIABLES AND PRELIMINARY CHOICES

During this thesis project, a dataset provided directly by Salvatore Ferragamo S.p.A. was used, containing real data related to three consecutive years of business activity (2022, 2023, and 2024), with a specific focus on the Handbags merchandise category. Access to this type of information represents an extremely relevant opportunity from both an academic and applied perspective: it allows for the construction of an experimental project fully embedded in the operational reality of a luxury maison, where decision-making processes, and time-to-market management, are based on concrete elements and not theoretical assumptions. Ferragamo examines this data and then discusses future product decisions with the regions in which it operates, each of which has specific needs within its reference market.

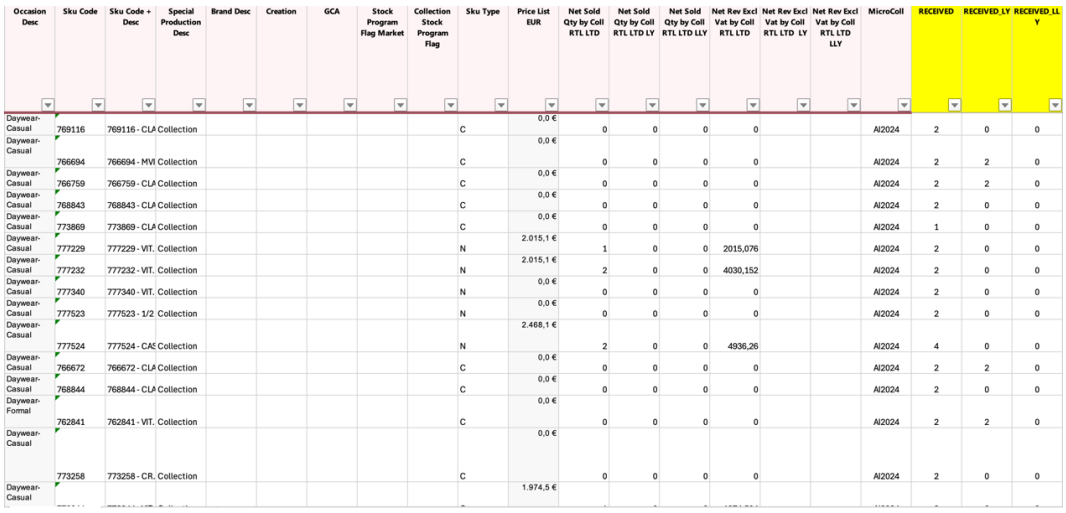
The dataset used is the result of a careful selection phase shared with the corporate team. The initial focus was placed on the Handbags family, because, as emphasized in several sessions with the Merchandising Planning managers, it represents one of the key categories for Ferragamo in terms of visibility, product turnover, and strategic relevance in building the brand image. This selection is not the result of a merely quantitative criterion but reflects a methodological logic, based on the idea that starting with a category with a high data density and relatively stable characteristics is the most suitable solution to test the robustness and scalability of the predictive algorithm. In other words, the handbag category represents an ideal laboratory, sufficiently broad to build meaningful models but at the same time circumscribed enough to maintain a high level of control over the variables analyzed.

The dataset includes data from all regions where Ferragamo is currently active, thus covering the seven strategic macro-areas of the group (Europe, United States, Japan, SEAP, LATAM, China, and Korea). This aspect is fundamental, as

it allows for reflections not only on a global level but also on a granular level of individual store or individual region, mapping in detail the behavior of SKUs in relation to the cultural, commercial, and logistical specificities of the different target markets. Example:



Region Desc	Country	Channel Desc	Store_code	Store Code Desc	Entity Typology	Sku Birth Collection	Prod Cat Gender Desc	ROTB Macro Category Desc	Production Category Code	Prod Category Code + Desc	Macro Merc Typology Desc	Merchandise Typology Desc	Macro Line Desc	Line Desc	Model Code	Model Name	Color Desc	Material Macro Group Desc	Material Group Desc	Dimension Group Desc
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	FERRAGAMO HUG	HUG SOFT	215608	HUG TH S	NYLUND PINK	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Clutch	Clutch	FERRAGAMO HUG	HUG CROCO PRINT	215890	HUG POUCH L	BLACK	Leather	CALF	Large
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	FERRAGAMO HUG	HUG SOFT	215608	HUG TH S	OPTIC WHITE	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Clutch	Clutch	FERRAGAMO HUG	HUG SOFT	215974	HUG POUCH S	LUCKY CHARME	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	FERRAGAMO HUG	HUG SOFT	215975	HUG MINI	AZUR	Leather	CALF	Mini
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Shoulder Bag	Flap	NEW LINE AI24	NEW LINE AI24	219794	FL M	NERO	Leather	CALF	Medium
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Shoulder Bag	Flap	NEW LINE AI24	NEW LINE AI24	219794	FL M	OXBLOOD	Leather	CALF	Medium
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Tote	Tote - Handheld	NEW LINE AI24	NEW LINE AI24	219805	TOTE L	NEW OLIVE	Leather	CALF	Large
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	FERRAGAMO HUG	HUG GRAINY	215608	HUG TH S	NEW OLIVE	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	FERRAGAMO HUG	HUG GRAINY BICOLOR	215608	HUG TH S	TESTA DI M	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Clutch	Clutch	FERRAGAMO HUG	HUG SOFT	215890	HUG POUCH L	YELLOW	Leather	CALF	Large
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Clutch	Clutch	FERRAGAMO HUG	HUG SOFT	215974	HUG POUCH S	NYLUND PINK	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	FERRAGAMO CREATION	F.CREATION	212193	TOP HANDLE S	LAPIS	Leather	CALF	Small
U.S.A.	Canada	Primary	U054	U054 - Vanco	Store		Women	Handbags	21	21 - Handbag	Top Handle	Top Handle	THE STUDIO	THE STUDIO BOX VENUS PRINT	211496	ST.BOX MINI	OPTIC WHITE/AZUR 211496	Leather	CALF	Mini
											Top Handle	Top Handle	THE STUDIO	THE STUDIO	211424	ST.BOX MINI	AZUR	Leather	CALF	Mini



Occasion Desc	Sku Code	Sku Code + Desc	Special Production Desc	Brand Desc	Creation	GCA	Stock Program Flag Market	Collection Stock Program Flag	Sku Type	Price List EUR	Net Sold Qty by Coll RTL LTD	Net Sold Qty by Coll RTL LTD LY	Net Sold Qty by Coll RTL LTD LLY	Net Rev Excl Vat by Coll RTL LTD	Net Rev Excl Vat by Coll RTL LTD LY	Net Rev Excl Vat by Coll RTL LTD LLY	MicroColl	RECEIVED	RECEIVED_LY	RECEIVED_LLY
Daywear-Casual	769116	769116 - CLA	Collection						C	0,0 €	0	0	0	0			AI2024	2	0	0
Daywear-Casual	766694	766694 - MVI	Collection						C	0,0 €	0	0	0	0			AI2024	2	2	0
Daywear-Casual	766759	766759 - CLA	Collection						C	0,0 €	0	0	0	0			AI2024	2	2	0
Daywear-Casual	768843	768843 - CLA	Collection						C	0,0 €	0	0	0	0			AI2024	2	0	0
Daywear-Casual	773869	773869 - CLA	Collection						C	0,0 €	0	0	0	0			AI2024	1	0	0
Daywear-Casual	777229	777229 - VIT.	Collection						N	2.015,1 €	1	0	0	2015,076			AI2024	2	0	0
Daywear-Casual	777232	777232 - VIT.	Collection						N	2.015,1 €	2	0	0	4030,152			AI2024	2	0	0
Daywear-Casual	777340	777340 - VIT.	Collection						N	0,0 €	0	0	0	0			AI2024	2	0	0
Daywear-Casual	777523	777523 - 1/2	Collection						N	0,0 €	0	0	0	0			AI2024	2	0	0
Daywear-Casual	777524	777524 - CAS	Collection						N	2.468,1 €	2	0	0	4936,26			AI2024	4	0	0
Daywear-Casual	766672	766672 - CLA	Collection						C	0,0 €	0	0	0	0			AI2024	2	2	0
Daywear-Casual	768844	768844 - CLA	Collection						C	0,0 €	0	0	0	0			AI2024	2	0	0
Daywear-Fomal	762841	762841 - VIT.	Collection						C	0,0 €	0	0	0	0			AI2024	2	2	0
Daywear-Casual	773258	773258 - CR.	Collection						C	0,0 €	0	0	0	0			AI2024	2	0	0
Daywear-Casual										1.974,5 €										

All stores in the dataset have already been divided into groups through a clustering process implemented by the company. Each store is logically located in different areas concerning its reference region, for example, city centers, shopping malls, airports, etc., and above all, each store has a different catchment area and surface area (m^2), which allows for a maximum number of SKUs per store. Precisely for this reason, the clustering process makes it possible to identify how many SKUs can be requested by the store for

each category. The problem, as previously explained, remains: how do we understand which SKUs should be selected?

The structure of the dataset was built in such a way as to combine product registry information, operational data, and performance metrics in an integrated logic. Among the main informative components are: Geographical and commercial identifiers: each row of the dataset is associated with a specific store, for which the store code (store_code), country (country_desc), and region (region_desc) information is reported. These fields allow the construction of regional clusters, the execution of comparative analyses between markets, and the analysis of product distribution in a way consistent with Ferragamo's real commercial segmentation. Product information: each SKU is described with several distinctive attributes including the SKU code, the associated product line, material, color, macro-category, and price range (e.g., "Low," "Medium," "High"). This information is essential for building predictive models that consider not only historical performance but also the intrinsic characteristics of the product. Moreover, it is worth noting that these are the only pieces of information available before a product is launched on the market. Sales and receiving metrics: the operational core of the dataset is represented by quantitative variables that measure the behavior of SKUs in different stores, including: received: number of units received for a given SKU in a specific year; net_sold_qty_by_coll_rtl_ltd: number of net units sold; sell_through: ratio between sold and received, a synthetic indicator of the distributional effectiveness and attractiveness of the SKU; temporal variables for LY (last year) and LLY (last last year), allowing for retrospective comparisons and performance analysis over extended time periods.

Structurally, the dataset is organized in long table format, meaning each row represents a specific combination of SKU, store, year, and product characteristics, with all associated variables reported in columns. This structure is particularly useful for aggregation operations, grouping, and statistical visualization, as well as for implementing machine learning models, which benefit from a clear separation between observations and features. The

dataset used can be described as a multidimensional database, sufficiently rich and detailed to support an experimental predictive analysis. The selection of handbags as a pilot category is consistent with both the company's needs and the technical constraints of the project, and the future extension of the model to other categories is already included in the project roadmap, making this initial phase a key step toward a broader data-driven transformation of Ferragamo's commercial planning.

3.2 DATA PREPROCESSING AND CLEANING: INITIAL ISSUES FOUND IN THE DATASET

The exploratory analysis conducted on the dataset provided by Ferragamo highlighted a series of structural and qualitative issues that became evident from the very first stages of loading and visualizing the data. To analyze the dataset, initially provided by the company as a .xlsx file in Microsoft Excel, Python was chosen due to its extensive libraries for statistical models and machine learning. In fact, every single step for the modulation and data cleaning process was carried out in Python. Returning to the dataset's problems, which, although partly physiological in any real and corporate context, require a very careful preliminary cleaning phase, without which any modeling attempt would be compromised by distortions, inconsistencies, or lack of generalizability.

One of the first significant aspects that emerged was the considerable presence of null values (NaN) within multiple variables. This lack of data was distributed heterogeneously throughout the dataset: in some cases, it concerned individual rows or specific stores; in other cases, entire columns displayed very low levels of completeness. In particular, the 'received' variable, one of the most important in the context of this study, presented numerous instances of missing data in 2024, which could be interpreted differently depending on the context. In some scenarios, a zero in 'received' could indeed indicate the absence of SKU shipment to the store in question, while in others, it might simply be a data point that was not entered or not correctly tracked in the system. The same applies to the

'net_sold_qty_by_coll_rtl_ltd' variable, which on more than one occasion shows null values even in active stores, raising doubts about the completeness of the information.

Therefore, the handling of NaNs cannot be approached in an automatic and undifferentiated manner: it requires a careful analysis of the underlying business logic for each field and, where possible, a direct consultation with internal teams to verify the actual validity of the data. In this case, using the pandas and numpy libraries in Python, it was possible to remove rows where the missing data could not be recovered or justified. For the year 2024, however, in agreement with the corporate team, it was decided to replace NaN values in the 'received' column with zeros.

A second major issue is related to the excessive complexity of the product description component, meaning the very high number of columns dedicated to encoding the physical and stylistic characteristics of each SKU. The dataset includes a substantial number of textual variables that, although theoretically meaningful (such as color, material, line, description, macro-category, price range), generate a very high level of informational dispersion. In some cases, the same variable is expressed in different ways across stores or years, resulting in duplicates, inconsistent formats, or virtually unique combinations that reduce the generalizability of models. The 'color' variable, for example, appears with numerous nearly equivalent labels written in different ways ("Nero", "black", "nero opaco", "black matte"), which, although describing the same concept, end up being treated as distinct categories. This creates a high-cardinality problem for variables that should ideally have a limited and manageable number of categories.

A similar problem is observed in the encoding of product lines and materials, where the high semantic variability, often not accompanied by centralized corporate documentation, makes it difficult to group the information coherently. This hyper-segmentation not only risks producing excessively complex models but may also obscure relationships between similar SKUs that, commercially, belong to the same family but are technically categorized

in divergent ways. To build a robust and generalizable model, it is therefore necessary to proceed with the simplification and rationalization of categorical variable encoding, through operations such as lexical standardization and merging of redundant categories. To resolve these issues, synthetic variables will be created.

An additional issue to note concerns the presence of duplicated SKUs or, more precisely, the recurrence of the same SKU code in combination with different temporal, geographic, or commercial attributes. This phenomenon is entirely normal in a multi-regional and multi-seasonal retail context, but it demands careful management of the analysis granularity level: it is necessary to decide whether to treat each row as a distinct instance (SKU-store-year), or to aggregate the data. Without a clear decision on this point, there is a risk of generating distortions in the calculation of means, variances, and performances, compromising the consistency of the results.

Lastly, the need to integrate new variables that can overcome the limitations of traditional metrics should not be underestimated. In summary, while the original dataset is of great value and relevance, it presents a series of challenges typical of real-world contexts, which must be addressed with a rigorous and systematic approach. Therefore, the cleaning phase should not be considered a secondary preliminary operation but rather one of the central stages of the project, upon which the credibility and effectiveness of any subsequent predictive analysis rest.

3.2.3 PREPARATION AND TRANSFORMATION OF THE DATASET: DATA CLEANING PROCESS IN PYTHON ENVIRONMENT

Following the initial exploratory phase described in the previous paragraph, it became necessary to intervene in a deep and systematic way on the dataset received from the company. As previously discussed, the file provided by Ferragamo presented numerous analytical potentials but, at the same time, revealed a series of structural and operational limitations that could have

compromised the quality of the predictive models to be built. For this reason, the transformation of the dataset was approached as an autonomous and articulated project phase, rather than as a simple technical step. All the work was carried out in a Jupyter Notebook environment, using the Python language as the primary tool for data manipulation and cleaning.

The entire process required an initial configuration of the working environment, starting from the installation and import of the fundamental libraries. The core of the manipulation was managed using the panda's library, widely recognized for its effectiveness in handling structured data in table format. In addition, numpy was used for mathematical operations on arrays and null values, matplotlib. pyplot for exploratory visualizations, and seaborn for generating heatmaps, boxplots, and scatter matrices useful for identifying outliers and distribution patterns. The libraries were imported with the following commands:

import pandas as pd, import numpy as np, import matplotlib. pyplot as plt ,import seaborn as sns

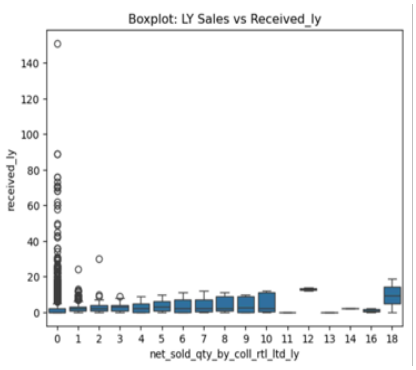
This phase was followed by the loading of the Excel file containing the complete dataset, named "db_handbags_22-23-24.xlsx", with particular attention to the specific worksheet containing the data. The loading was managed using the pandas read_excel () method. To make the content of the file readable and understandable from the outset, the first records were displayed using the. head() method and the number of rows and columns was examined, along with the presence of null values, using df.info() and df.isnull().sum().

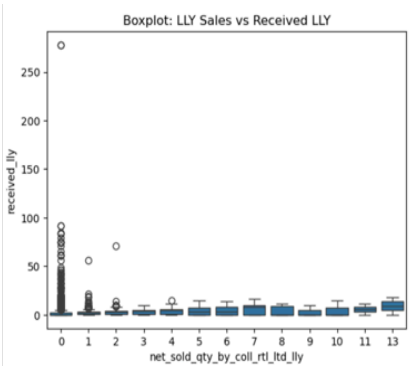
At this point, a critical issue emerged: the massive presence of NaN values distributed across key variables. In particular, the columns 'received' and 'net_sold_qty_by_coll_rtl_ltd' were affected by a significant share of missing data. In addition to these quantitative columns, the 'GCA' and 'Stock Program' columns had no data at all, although this was due to company privacy restrictions. This phenomenon required interpretative reflection: in

fact, a missing value does not always indicate an omission or registration error. In some cases, a null value in 'received' could indicate an actual lack of physical product receipt in the store. However, without structural confirmation from the company's management systems (SAP and Board), it was impossible to clearly distinguish between an error and a missing data point due to real causes. The company therefore decided to solve this problem by replacing NaN values in 'received' for 2024 with zero.

Moreover, some 'received' values for 2024 turned out to be negative, an impossible value caused by a different programming language used in Board, which interprets negative values as returns from customers. To preserve the robustness of the analysis, it was decided to exclude all SKUs that did not have at least one valid receipt, i.e., with a 'received' value greater than or equal to 0.

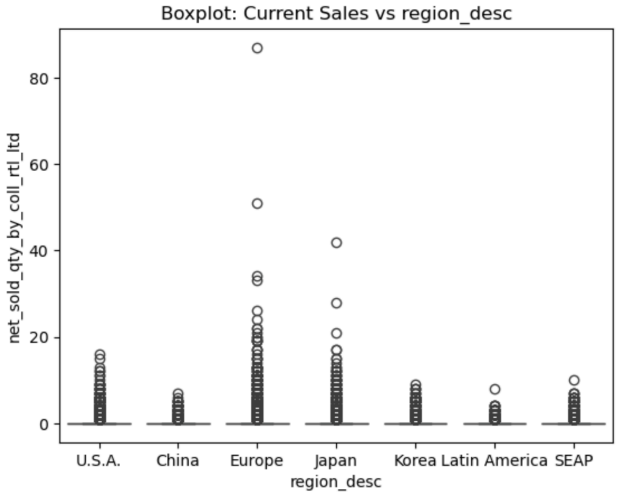
This issue also emerged during the analysis of 'sell_tr' over the various years, which displayed anomalous values either greater than or less than 100%. Subsequent checks were performed to examine the remaining rows. A count revealed 68,871 SKU-Country-Store combinations remaining out of an initial 84,000, representing a significant cleaning. The initial processes revealed that there were many more 'received' values in 2022 and 2023 compared to 2024, where the data was incomplete due to some SKUs being under analysis by Ferragamo and protected by company privacy. In fact, the count of 'received_ly' and 'received_lly' exceeds 100 received products in some stores.





As can be seen from the graphs above, very few sales were recorded. One of the main causes was undoubtedly the post-Covid crisis that affected, as

mentioned in the early chapters, a large part of the fashion industry. According to company sources, the 'received' figures recorded in those years were very similar to the pre-Covid values, but the post-pandemic crisis drastically halved sales.



Here, however, are the sales recorded in the first period of 2024, which show a marked recovery, despite having significantly lower average 'received' values, due to Ferragamo's much more cautious approach.

A second area of intervention concerned the standardization of categorical values, particularly those related to product description: color, material, line, and macro-category. These variables were formally separate but conceptually overlapping and exhibited notable orthographic and semantic variability. For example, the color black was represented in at least six different ways: "Black", "BLACK", "black matter", "nero", "NERO", "nero opaco". This type of redundancy results in an unnecessary multiplication of categories and makes any predictive or descriptive logic based on semantics inapplicable.

To address the problem, a series of normalization functions were applied to convert all strings to lowercase (.str. lower()) and remove whitespace and special characters (.str.strip() and str.replace()). In addition, some features were grouped. For the 'colour_group' variable, a 'Macro_colour_group' variable was created, grouping colors by hue into a chromatic scale, reducing the number of dummy variables from 35 to 8.

Example:

df_clean['color'] = df_clean['color']. replace ({ 'nero opaco': 'black', 'nero': 'black', 'black matte': 'black', 'black ': 'black', 'black.': 'black' })

The same logic was applied to the 'Material', 'Line', and 'Macro-category' variables, always trying to identify homogeneous logical groups and reduce the cardinality of categorical variables. It was essential to prevent the number of categories from exceeding the threshold beyond which machine learning algorithms (such as Poisson Regression) lose efficiency or assign random weights to poorly represented categories.

The next crucial step concerned the management of duplicates and the definition of the observation granularity. Each row of the dataset represented a combination of SKU, store, year, and physical characteristics, but in many cases, the same SKU appeared in multiple rows due to its presence across several regions or years. The central question posed by the project was therefore: what analytical unit should be used for prediction? It was decided to maintain only 2024. presented an excessive number of zero values in the sold_qty variable. These zero entries introduced a significant imbalance in the dataset, due to post-covid market period, potentially distorting the model's ability to learn meaningful patterns. By focusing exclusively on 2024, which reflects a more balanced year to analyze for the handbags segment, the analysis remains more robust and representative of current market dynamics

Throughout the entire cleaning phase, a logic of continuous control over the structure of the dataset was maintained, using commands such as df.describe(include='all'), df.dtypes, and df.nunique(). This approach made it possible to detect potential problems in real time, such as the presence of extreme values, non-numeric variables erroneously interpreted as objects, or completely empty columns to be removed.

A final reflection concerns the documentation of operations. Each step was documented within the notebook using detailed comments in natural

language, thus ensuring the reproducibility of the process and transparency of the choices made. In the context of the thesis, this also represents a formative opportunity: learning to explain each line of code not as a mere “script” but as a methodological act justified by empirical needs and analytical objectives.

CHAPTER 4- PREDICTIVE MODEL

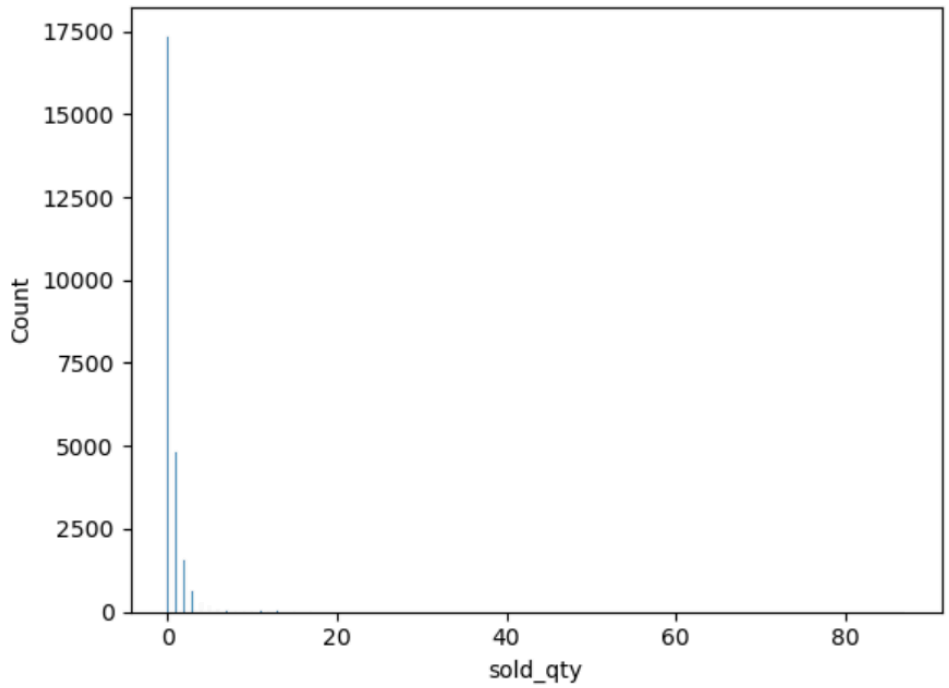
4.1 Introduction to the post-cleaning exploratory phase

After completing a complex and rigorous data cleaning process, already illustrated in the previous chapters, an analytical exploration process was launched aimed at gaining a detailed understanding of the behavior of the target variable sold_qty, namely the number of units sold for each SKU within the 2024 collection, with the goal of establishing which variables had an impact on sold_qty and did not affect the model. This phase represents an essential step within the modeling process, as it allows the identification of significant patterns, potential structural anomalies, and non-trivial relationships with the explanatory features.

The variable sold_qty is a discrete numerical variable and plays a central role in this thesis. It represents not only the final target to be predicted but also a synthetic indicator of the commercial performance of the individual SKU within each combination of store and region. The main objective will therefore be to identify which product characteristics and geographic context features are most strongly correlated with the variation of sold_qty, thus providing the basis for a future predictive and optimized allocation of SKUs at the regional level.

The first operation performed in the notebook PROVA_MODELLI was the exploration of the univariate distribution of the variable sold_qty, through a frequency histogram. The chart clearly shows a highly skewed distribution: the majority of SKUs display very low values (0 or 1 units sold), while only a minority reaches values above 5 units, a value that was expected. For this reason, during the cleaning phase it was decided to narrow down these values by counting only 2024. A broader discussion could be opened here, considering that zero values still represent a form of data – negative, yes, but still a data point indicating that a given SKU did not sell. However, in this case, a strategy previously adopted by the company, which was deemed a failure , had already been implemented, namely continuing to bet on the same production numbers even in the post-COVID period. This means that such data could skew predictions in the model. As previously stated, the data

collection and its layout are not adequate to understand what specific causes led to this failure, which is why a more restricted vision was preferred.



Following these first observations, the total number of observations was calculated, equal to 25,117 rows, each representing a unique combination of SKU, store, and geographic region. The unbalanced nature of the dataset motivated the adoption of appropriate modeling approaches, which will be discussed in the following paragraphs.

4.1.1 Study by region: heterogeneity in geographical performance

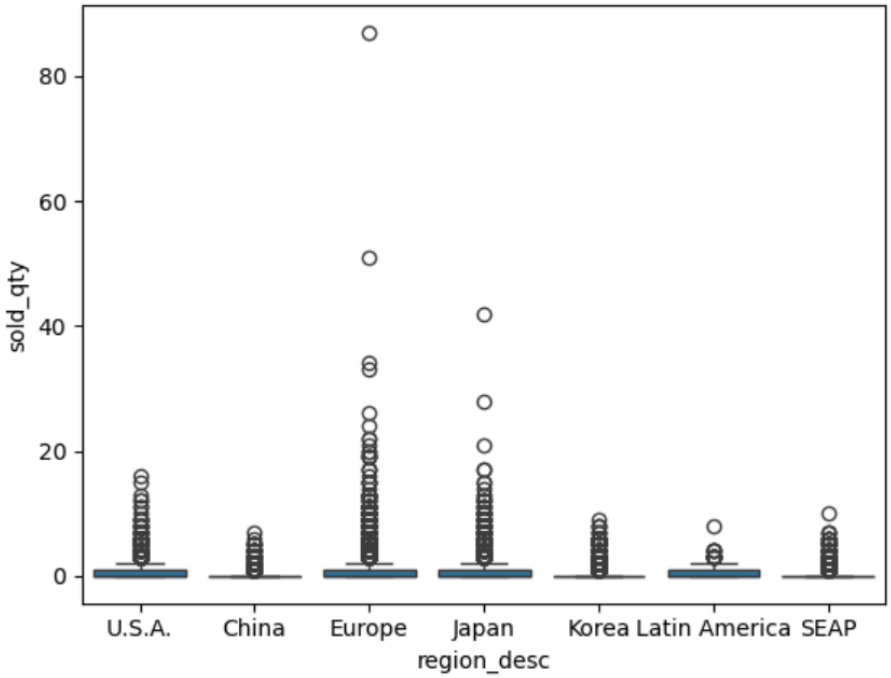
After the visualization of the boxplot, a tabular view of the values present in the dataset was then produced:

•The number of observations (n_obs),

•The mean and median of sales,

•The standard deviation,

•The percentage of SKUs that sold at least one unit.



	region_desc	n_obs	sold_qty_mean	sold_qty_median	sold_qty_std	sold_qty_positive_pct
0	Europe	4196	1.174690	0.0	2.721869	0.486892
1	Japan	3017	0.951608	0.0	1.954636	0.432217
2	U.S.A.	4724	0.671677	0.0	1.208321	0.385478
3	Latin America	1551	0.386202	0.0	0.685172	0.303030
4	SEAP	3517	0.352005	0.0	0.805664	0.231447
5	Korea	3530	0.324646	0.0	0.849042	0.196601
6	China	4582	0.178306	0.0	0.532572	0.134439

The results confirm a significant geographical heterogeneity. In addition to a macro view of the Regions, the code also explored the more detailed part of each individual Region, the Countries.

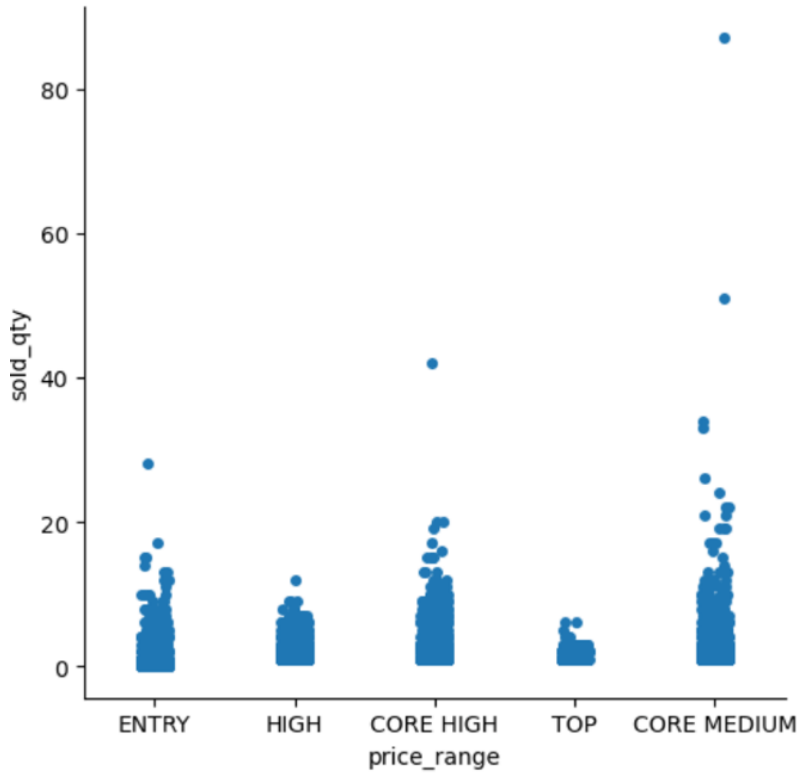
The Countries were a focal point in this analysis, since the stores present in each country are at the center of the analyses already carried out by Ferragamo to assign a cluster to each store, in the dataset store_code. Through this cluster, as already explained, the number of SKUs that can go inside it is determined, as well as the store visualization cycle that takes place. Each store has different sales potential, visible within the dataset, and useful to the model for detecting the various changes depending on the reference store. Region and Country are the two variables outside the descriptive characteristics of the product. These are variables that are useful for planners and buyers in determining which SKUs to prioritize during the purchase phase.

4.1.1 Analysis of the main explanatory variables

The notebook then proceeds to visually explore the relationship between sold_qty and some features within the dataset, mainly categorical. For each variable, analyses are performed in response to sold_qty, to determine which features are useful for building a model. Through the use of boxplots and tabular values, the aim is to explore each feature.

The features analyzed:

Price_range: The first feature analyzed is the price range, and it is evident that sales are well distributed across different types of price ranges. Maximum sales were reached by SKUs in a medium price range, but, as later shown in tabular format, the number of selling SKUs reveals that “HIGH” has a total number of SKUs with sold_qty > 0 higher than that of the medium range. All of this is also determined by the number of received, already explored earlier. In summary, price_range proves to be a determining feature for the structure of the model.

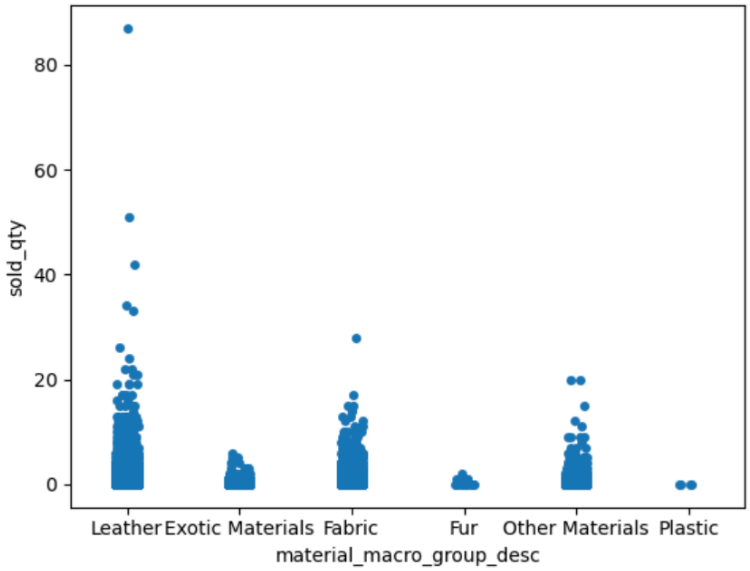


	price_range	n_sku_sold_positive_price
0	HIGH	267
1	CORE HIGH	234
2	CORE MEDIUM	140
3	TOP	111
4	ENTRY	66

Colour_group: This will not be analyzed, since during the data preparation phase a Macro_colour_Group was created, based on color shading, due to the strong risk of constantly introducing new unique colors and ending up with too many one-off variables. Color is one of the few physical characteristics available before production. There are annual trend studies, so the model could help in designing the SKUs and understanding which color might perform better depending on the market. In the target variable, it is uniformly distributed, indicating that the macro_colour_group will not create redundancy and will be useful for model structure.

occasion_desc: This variable presents only a few values (3), which are not influential in terms of sold_qty.

material_macro_group_desc: This feature was thoroughly analyzed, especially due to a reasoning similar to that of color, meaning it is a characteristic that could strongly aid product development. However, here a significant gap emerges regarding the category treated in this thesis, handbags, a category that heavily favors the use of leather. As shown in the table, the majority of SKUs are sold in leather. Tests were conducted on including this feature, confirming what was previously stated: the deviance in the GLM model increased with critical values. A possible improvement, since this remains a highly exploratory feature, would be to divide the materials into even more macro groups, but such a division would only be possible with Ferragamo's know-how. The same applies to the insertion of micro-groups for materials, where "calf" would distort the training phase too much.



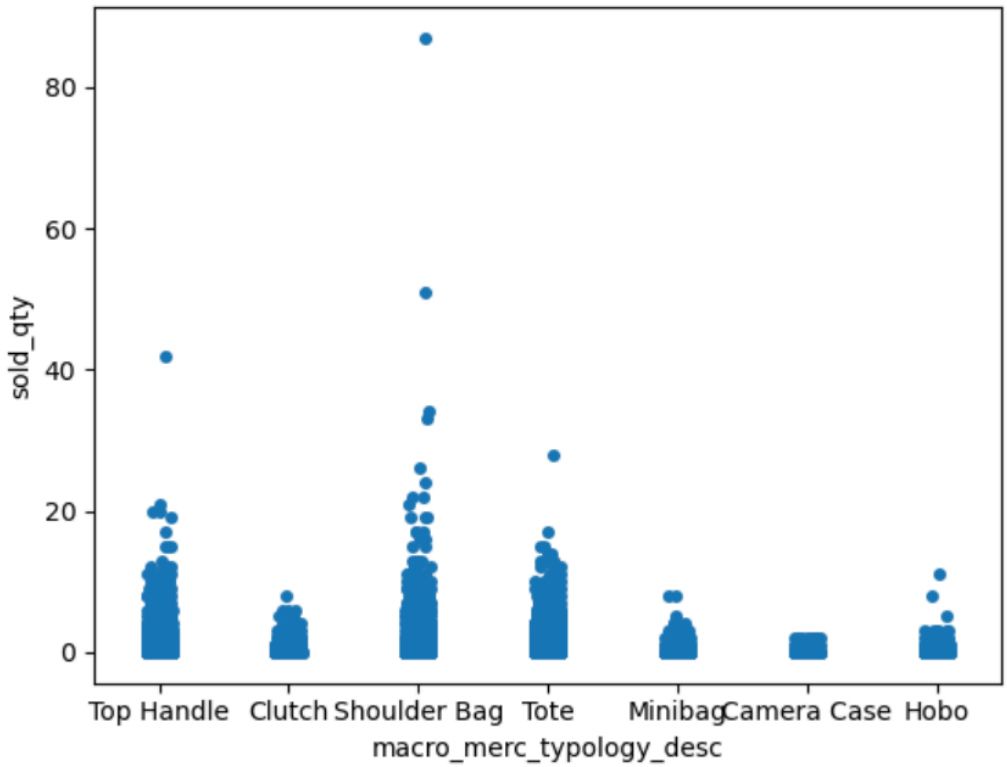
dimension_group_desc: Another important product characteristic that makes it possible to identify and analyze performance across different markets. This feature has 4 variables, with a well-distributed and excellent response in terms of sold_qty, indicating it is a highly influential variable for sales and product performance.

	dimension_group_desc	n_sku_sold_positive_dimension_group_desc
0	Small	207
1	Medium	101
2	Large	59
3	Mini	55

macro_merc_typology_desc: The merchandise typology describes the type of handbag, which is important in exploring sold_qty. Within the notebook, a regrouping is carried out as some rows had unique characters.

```
sns.stripplot(data=df, x='macro_merc_typology_desc', y='sold_qty')
```

```
<Axes: xlabel='macro_merc_typology_desc', ylabel='sold_qty'>
```



macro_line_grouped: The analysis of macro lines was conducted differently. This feature responds very well, increasing the levels of R-squared, but again, the same problem seen with color arises: the duration of the variables. Each year, Ferragamo creates new macro product lines when launching collections, which is a significant obstacle for the model, as it would not know how to classify the new variables or predict their performance. So, the same point discussed earlier applies: a longer-lasting division over time would be better. In this case, however, the macro line is also considered a descriptor of the type of bag being analyzed, whether it features materials like “gancini” or a unique shape like the “clutch”.

In the notebook, it is noted that the features region and country contain many values with 0 and could create collinearity, but thanks to the many variables within the above-mentioned features, it is still possible to read different situations and to calculate performance.

For the model, the following features were therefore used:

‘region_desc’,

‘country’,

‘price_range’,

‘macro_colour_group’,

‘macro_merc_typology_desc’,

‘dimension_group_desc’,

‘macro_line_grouped’,

‘store_code’

Excluding many features such as channel_desc, entity_typology, … received and sell_tr, the latter being a consequence of sales and therefore not appropriate for an analysis of new products.

This last point should be seriously considered, as the main objective is to predict the performance of new SKUs. The selected features are characteristics known before the launch and on which it is possible to work to find the best combination, a very important aspect.

4.1.2 First modeling approach: binary classification

Within the developed analytical framework, one of the main challenges consisted in classifying each SKU into one of two binary categories: “vendente” (1) or “non vendente” (0).

This type of classification represents a clear business objective: to support buyers and planners in the preventive identification of items with the highest

probability of generating actual sales, thus guiding assortment decisions and optimizing the Global Core Assortment.

After defining the target variable as sold_qty > 0, only categorical variables (region, store, price range, line, color, size) were selected to avoid ambiguities due to potentially unstable numerical values.
The predictor matrix X was prepared using one-hot encoding, while the target y represented the presence or absence of sales.

To proceed with the estimation, the dataset was split into two subsets:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

With this instruction, 80% of the data was allocated to the training set (used to train the model), while the remaining 20% was reserved for the test set, useful for evaluating out-of-sample performance.
The use of the parameter random_state=42 ensures reproducibility of the results by fixing the randomness seed.

The model chosen for classification was the Logistic Regression Classifier, an initial choice consistent with the need for interpretability, ease of implementation, and a first validation of the predictive effectiveness of categorical variables alone:

```
clf = LogisticRegression(max_iter=1000)
clf.fit(X_train, y_train)
```

The algorithm learned the weights of the independent variables through the logistic (sigmoid) function, iterating up to a maximum of 1000 epochs to ensure convergence.
The use of logistic regression is particularly useful in this context as it allows the estimation of the probability that an SKU will be sold based on its characteristics.

Once trained, the model was used to perform two types of predictions:

```
y_pred = clf.predict(X_test)
y_proba = clf.predict_proba(X_test)[:, 1]
```

•y_pred: returns the predicted class for each observation (0 = non vendente, 1 = vendente)

•y_proba: returns the estimated probability that each SKU is indeed vendente.

This output is essential for ROC curve analysis, AUC calculation, and future applications in which the decision threshold may be dynamically calibrated.

This implementation represents the first step towards a complete classification pipeline.

The logistic model, although with some limitations due to the exclusive use of categorical variables and the absence of behavioral or quantitative data (such as sell-through, received, promo, etc.), has nevertheless proven capable of capturing significant patterns and representing a solid interpretive base for evaluating the probability of success of each product.

The classifier's performance was analyzed through indicators such as accuracy, recall, precision, ROC curve, and confusion matrix, discussed in detail in the following section.

The comparison between predictions and reality showed behavior consistent with the distribution of the target variable and made it possible to identify SKUs correctly predicted as 'vendente' even in the presence of suboptimal conditions (e.g., secondary stores, complex lines, atypical palettes).

After the training and validation phase of the logistic regression model, predictions were made on the entire test set, and then the main evaluation metrics were calculated:

accuracy, ROC-AUC curve, confusion matrix, and complete classification report (precision, recall, F1-score).

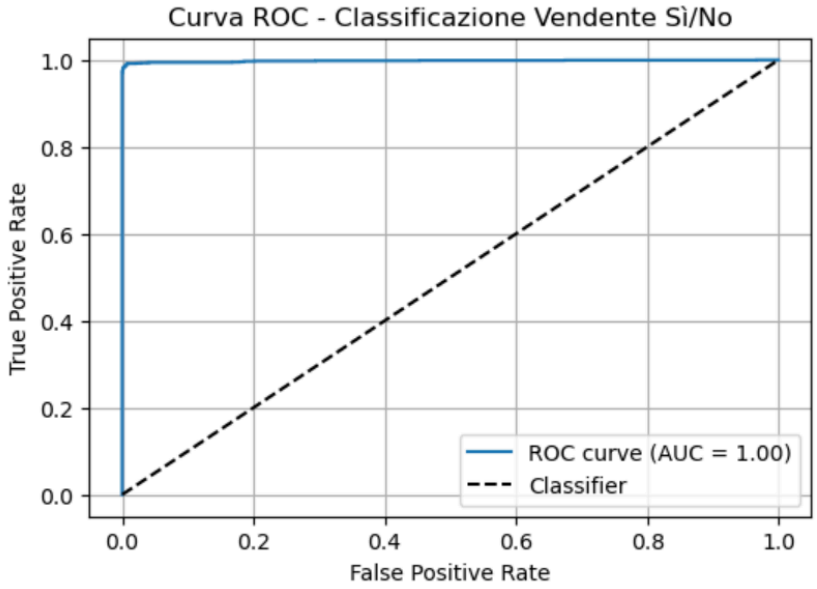
The results obtained are noteworthy, especially considering that the model is based solely on categorical variables and does not use numerical behavioral data.

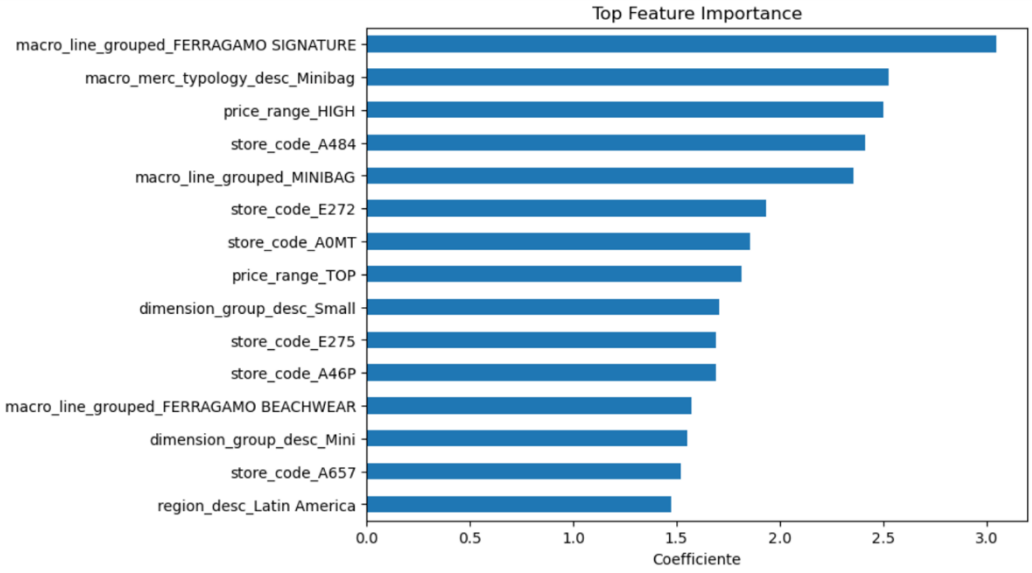
Accuracy and ROC AUC: an initial look at model quality

```
Accuracy: 0.9930334394904459
ROC AUC: 0.9977262191915045
Confusion Matrix:
 [[3465    8]
 [  27 1524]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      3473
           1       0.99      0.98      0.99      1551

    accuracy                           0.99      5024
   macro avg       0.99      0.99      0.99      5024
weighted avg       0.99      0.99      0.99      5024
```

The first indicator to consider is accuracy, equal to 99.3%, which represents the percentage of observations correctly classified by the model.
This already very high value is further reinforced by the AUC ROC (Area Under the Curve), equal to 0.9977, which quantifies the model's discriminative ability.

An AUC value so close to 1 indicates that the model is extremely effective in distinguishing between vendente and non vendente SKUs, even in the presence of variability between regions or imbalanced distribution characteristics.





Confusion Matrix: Detailed Interpretation

The confusion matrix returns the absolute number of correct and incorrect predictions for each class:

Class 0 (non-vendente):

•3,465 SKUs were correctly classified as non-sellers

•8 SKUs were incorrectly classified as sellers

Class 1 (vendente):

•1,524 SKUs correctly identified as sellers

•27 false negatives, that is, selling SKUs classified as non-sellers

Overall, the number of false negatives (27) and false positives (8) is extremely low relative to the total, confirming the stability and generalizability of the model even in a real scenario composed of numerous and heterogeneous SKUs.

Precision (class 1 = selling SKU): 0.99

•99% of SKUs classified as sellers are indeed sellers

•This is essential for the business, as it means the model does not systematically suggest false positives

Recall (class 1): 0.98

•98% of the SKUs that actually sold were correctly identified

•A value like this ensures that few selling SKUs are “lost,” which is crucial to avoid missed commercial opportunities

F1-score: 0.99

•The harmonic mean between precision and recall, very high, confirms the balance of the model.

A possible upgrade could involve identifying the sell_tr of certain SKUs from the previous year. However, there is not enough available data in the current database to support a proper exploration.

It should be emphasized that these results should not be interpreted negatively. On the contrary, the fact that a model built with only categorical variables can reasonably distinguish SKUs with high from low probability of being sold is already an important result.

This approach does not replace human decision-making, but supports it, providing a structured foundation to guide selection choices. In the future, model performance could be further improved by introducing balancing techniques, integrating additional variables available before launch, and testing more sophisticated architectures such as XGBoost or Gradient Boosted Trees.

An interesting extension could involve building separate models by product line or region, to further customize the predictions.

4.2 Poisson Regression Model: Preliminary Analysis and Diagnostics

4.2.1 Introduction to the Model Choice

In the phase following the descriptive study of the variables with respect to the target variable sold_qty and the classification model, it was decided to proceed with a Poisson Regression, consistent with the nature of the available dataset. The variable sold_qty represents the number of units sold for each observation (SKU-store), and is therefore a positive integer count with a strongly skewed distribution, as highlighted in the exploratory graphs.

The choice of the Poisson regression model stems from a precise methodological necessity: to adopt a predictive approach that respects the distribution of the response and can model the expected average sales rate as a function of qualitative characteristics of the product, the store, and the geographic region, without forcing normality assumptions, which would have been strongly violated.

However, in this specific modeling phase, numerical variables such as received or sell_through were not included, even though they would probably have improved predictive performance. This decision was intentional, in order to test the explanatory effectiveness of categorical variables alone within a basic Poisson formulation, ensuring the model could be used in pre-launch phases despite its known limitations.

4.2.2 Adopted Formula and Model Structure

At the heart of this analysis lies the construction of a predictive model of the Generalized Linear Model (GLM) type, with a log link function and Poisson probability distribution, aimed at estimating performance for each SKU.
Unlike the binary classifier presented in the previous section, which answers the question “will this SKU sell at least one unit?”, the Poisson regression aims to quantify the expected volume of sales, providing a continuous and interpretable estimate.

The model was applied to a dataset subset filtered to include only cases with sold_qty > 0 and < 20.
The GLM Poisson model was structured according to the following formula:

sold_qty~C(region_desc)+C(country)+C(price_range)+C(macro_colour_group) +C(macro_merc_typology_desc) + C(dimension_group_desc) + C(macro_line_grouped) + C(store_code)

This formula includes only categorical variables, all transformed into factors (C(...)) to allow the model to estimate the parameters of each level relative to a reference. Below is a detailed description of the features involved:
•region_desc: indicates the macro geographic area (e.g., Europe, Japan, China, USA)
•country: identifies the specific country of each store
•price_range: price range of the product (e.g., High, Medium, Entry)
•macro_colour_group: chromatic group of the product (e.g., Black, Beige, Multicolor)
•macro_merc_typology_desc: commercial category of the product
•dimension_group_desc: macro-cluster of the product’s size (Small, Medium, Large)
•macro_line_grouped: aggregated stylistic line (e.g., Studio, HUG, Archive)
•store_code: unique code of the store

The goal is to estimate how these characteristics influence the expected number of units sold and, above all, the performance for each product-store

combination, assuming that sold_qty is distributed according to a Poisson distribution.

The model was implemented using the Statsmodels library (statsmodels.formula.api) in Python, specifically the glm() function with the specification family=Poisson(). The fit() function allowed for coefficient estimation through maximum likelihood.

```
poisson_model = smf.glm(formula=formula,
                        data=df_poisson,
                        family=sm.families.Poisson()).fit()
```

With this code:

•The Poisson distribution family is specified, suitable for integer count variables

•The formula internally manages the encoding of categorical variables

•The dataset df_poisson contains only the real SKUs selected for the test, already pre-processed and cleaned

•The fit() method estimates the model's parameters, that is, the coefficients associated with each level of every categorical variable, in logarithmic terms of expected sales rates

After testing and validating the structure of the predictive model through simulated data and binary classifications, the final and most important phase of this research focused on applying the Poisson Generalized Linear Model (GLM) to a real dataset provided by Ferragamo.

This subset contains 7,750 observations related to handbag SKUs, actually sold or distributed in different stores and regions during previous seasons.

The central goal of this model was to reliably estimate the sold_qty variable, i.e., the number of units sold for each SKU, starting from a combination of categorical variables that represent:

•Geographic location (region, country)

•Product features (color, size, line, price range)

•Market characteristics (commercial typology, store typology)

This choice is fully consistent with the project's aim: the goal was not to estimate the exact sales figure (which can be influenced by random factors like ad hoc campaigns), but rather to model expected performance based on objective attributes of the product and context.
This is to support buyers in the early selection of SKUs, reduce the risk of unsold inventory, and optimize time-to-market.

Model quality: overall results

Generalized Linear Model Regression Results

Dep. Variable:	sold_qty	No. Observations:	7750
Model:	GLM	Df Residuals:	7432
Model Family:	Poisson	Df Model:	317
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-11679.
Date:	Tue, 16 Sep 2025	Deviance:	5077.9
Time:	01:40:02	Pearson chi2:	6.04e+03
No. Iterations:	100	Pseudo R-squ. (CS):	0.3053
Covariance Type:	nonrobust		

From a statistical perspective, the model achieves a Pseudo R-squared of 0.3053, a value that is not optimal but still noteworthy for a Poisson regression with over 300 dummy variables and strong interstore and interregion variability.
This value indicates that about 30% of the variance in sales levels is explained by product and context categorical features alone, without any numerical variable (such as received, sell_through, or moment), and without behavioral indicators.

Other key technical indicators:

•Deviance: 5,077.9

•Pearson Chi^2: 6,040.0

•Number of iterations: 100 (indicating good convergence stability)

These values confirm that the model is well-calibrated and, above all, robust, even in a context characterized by strong data heterogeneity.
Many of the independent variables shown in the coefficient table reported statistically significant values (p-value < 0.05), which suggests the presence of recurring patterns that can support SKU allocation decisions by geography and product typology. In particular:

Price range (price_range)

•TOP and HIGH SKUs show negative and significant coefficients.
•However, as previously mentioned, they still have more SKUs with sold_qty > 0
•On the contrary, ENTRY and CORE MEDIUM SKUs appear more versatile and adaptable

Color (macro_colour_group)

•The BROWN group has a strong positive impact on sales (coef. +0.279), while LILAC shows a significant negative effect
•Neutral palettes (NEUTRAL, RED_PINK) have moderately positive effects, confirming a preference for versatile tones in luxury fashion

Size (dimension_group_desc)

•Mini, Small, and Medium sizes are strongly positively correlated with sales volume
•In particular, Mini bags (coef. +0.649) are among the top performers
•This reflects a market trend, especially in Asian regions, toward compact, versatile formats

Stylistic line (macro_line_grouped)

•Lines like FERRAGAMO HUG (+0.497), SIGNATURE, STUDIO, and NEW LINE AI24 show highly significant and positive coefficients, confirming the success of the post-2022/2023 stylistic relaunch
•Historic lines like FIAMMA or capsule collections like ARCHIVE show more uncertain impact

Commercial typology (macro_merc_typology_desc)

•Tote bags (+0.363) and shoulder bags (+0.335) are more likely to sell, Clutch or hobo models show no significant impacts, suggesting that functional typology is a good performance predictor

Strategic interpretation of results

From a business perspective, these outputs demonstrate that forecasting SKU performance is possible even in the pre-launch phase, using only variables known upstream (design, materials, color, pricing, destination store). This can change the traditional operational paradigm.

This logic promotes greater global coherence (GCA), reduces the decision-making burden on buyers, and anticipates the creation of assortments more aligned with the local market, without compromising the aesthetic identity of the brand.

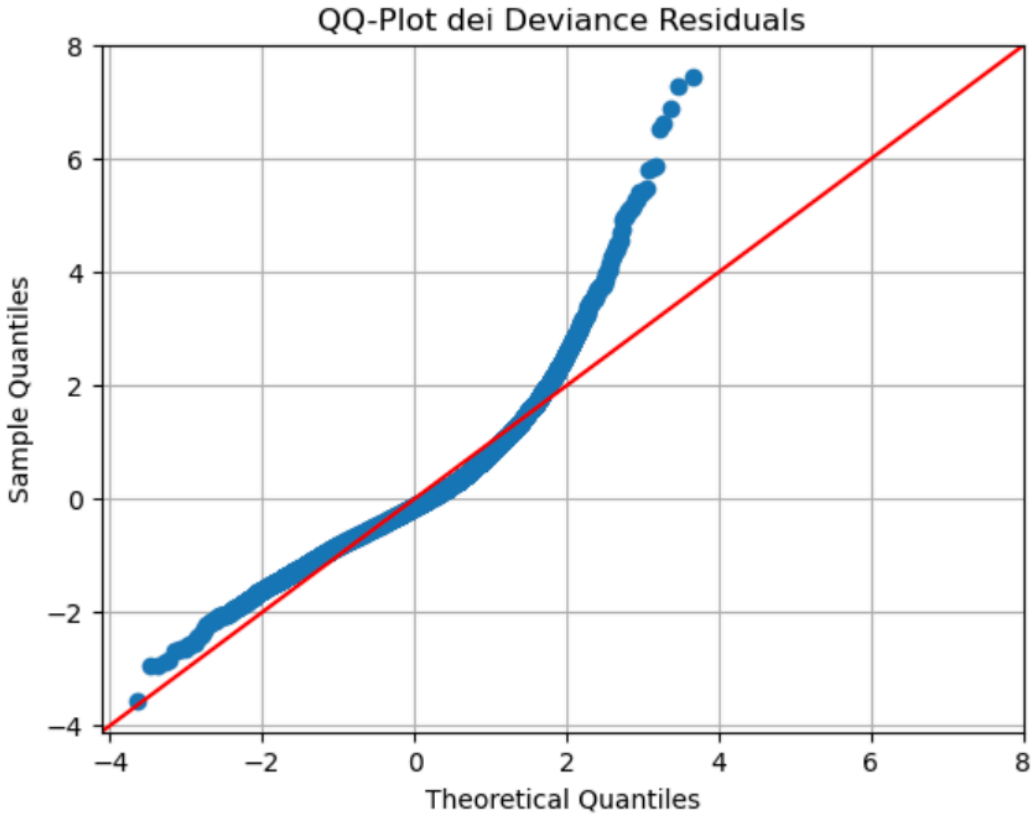
Limitations and possible improvements

Despite the excellent overall performance, it is important to highlight that:
•Many store_code and country variables are not statistically significant, and this is due to extreme fragmentation and the limited number of sales per single SKU-store.

•The model does not include numerical variables (such as received, sell_through, moment) by methodological choice, but including them in the future could increase accuracy and granularity.

•Some very high and unstable coefficients (e.g. >10^10) on certain dummies suggest a collinearity problem or categories with very few observations, a normal issue with many dummy variables.

QQ-Plot: Model Diagnostic



The graph shown represents a QQ-plot (Quantile-Quantile plot) of the deviance residuals from the Poisson regression model. This type of visualization is commonly used to assess the normality of residuals or, more generally, to compare the empirical distribution of residuals with the expected theoretical one.

What the graph shows:

•X-axis (Theoretical Quantiles): represents the theoretical quantiles of a standard normal distribution. If the residuals followed a perfect normal distribution, they would lie along a straight line.
•Y-axis (Sample Quantiles): shows the quantiles calculated from the deviance residuals obtained from the Poisson model.
•Red diagonal line (line='45'): is the reference line. It indicates perfect alignment between theoretical and observed quantiles. If all residual points lay on this line, it would mean the residuals follow the expected distribution.
•Blue dots: represent the actual residuals from the model. Each point corresponds to a deviance residual for a specific dataset observation.

Visual observations:
• Central behavior: in the middle of the plot (between about -2 and +2 on the theoretical quantiles axis), most points lie close to the red line. This indicates that the majority of residuals behave as expected, and the model is well-calibrated in this central range.
• Upper and lower tails:
• On the right side (theoretical quantiles > 2), there is a noticeable upward deviation: points diverge increasingly from the red line. This suggests that residuals in the upper tails (observations with much higher sold_qty than average) are larger than expected under the theoretical assumption.
• On the left side (quantiles < -2), a slight deviation is also observed, though less pronounced.

This tail behavior suggests that the model struggles to accurately capture the most extreme cases, especially for SKUs with high sales.
However, the generally consistent behavior in the central part of the graph shows that, for the majority of observations, the model provides a good estimate of expected values, and the residuals are well distributed around the mean.

Dispersion statistics
•Pearson Chi^2: 6042.39

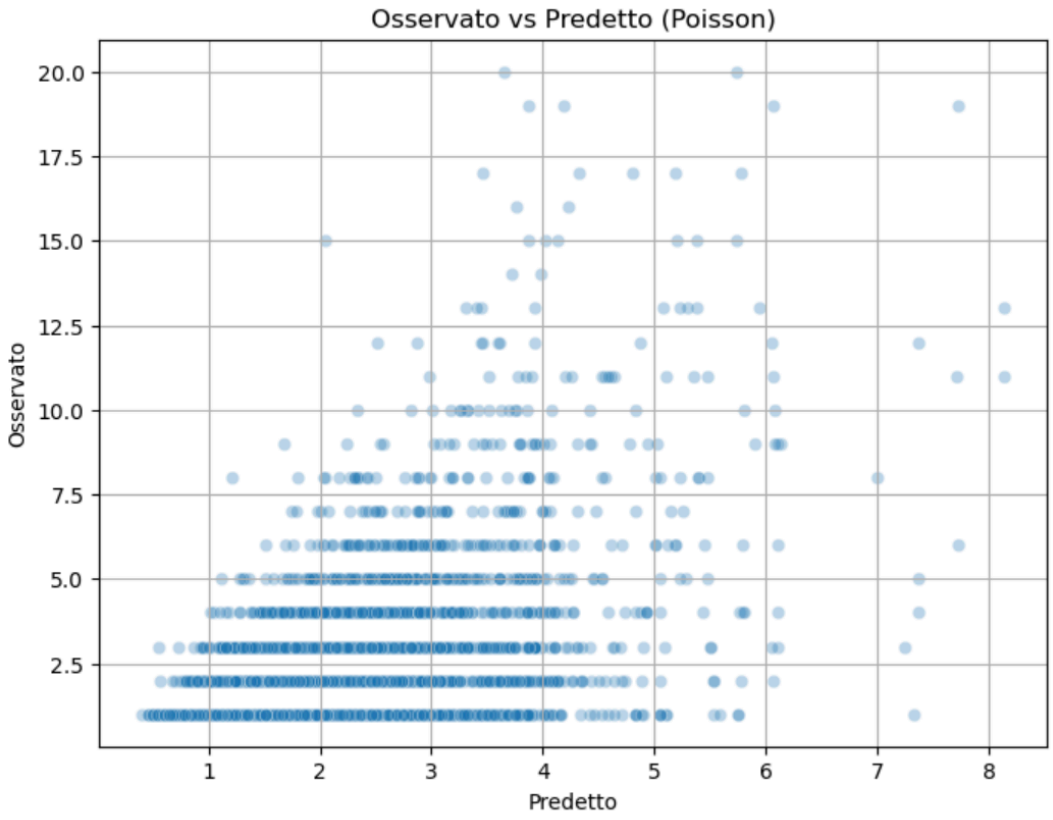
•Residual DF: 7432

•Dispersion parameter: 0.8130

The Pearson Chi^2 value of 6042.39, over 7432 degrees of freedom, yields a dispersion parameter of 0.81, which is less than 1. This indicates underdispersion in the Poisson model, meaning the observed variance in the data is slightly lower than the mean, in contrast to the classical equidispersion assumption (variance = mean).

In this case, the model is consistent and does not show signs of overdispersion, which would have required an alternative model (e.g., Negative Binomial).

Observed vs Predicted Plot



This graph visually compares the observed values with those predicted by the Poisson regression model. Each point represents a specific observation, i.e., a specific SKU in a given sales context.

•X-axis (Predicted): shows the sales quantities predicted by the model for each observation.

•Y-axis (Observed): shows the actual quantities sold.

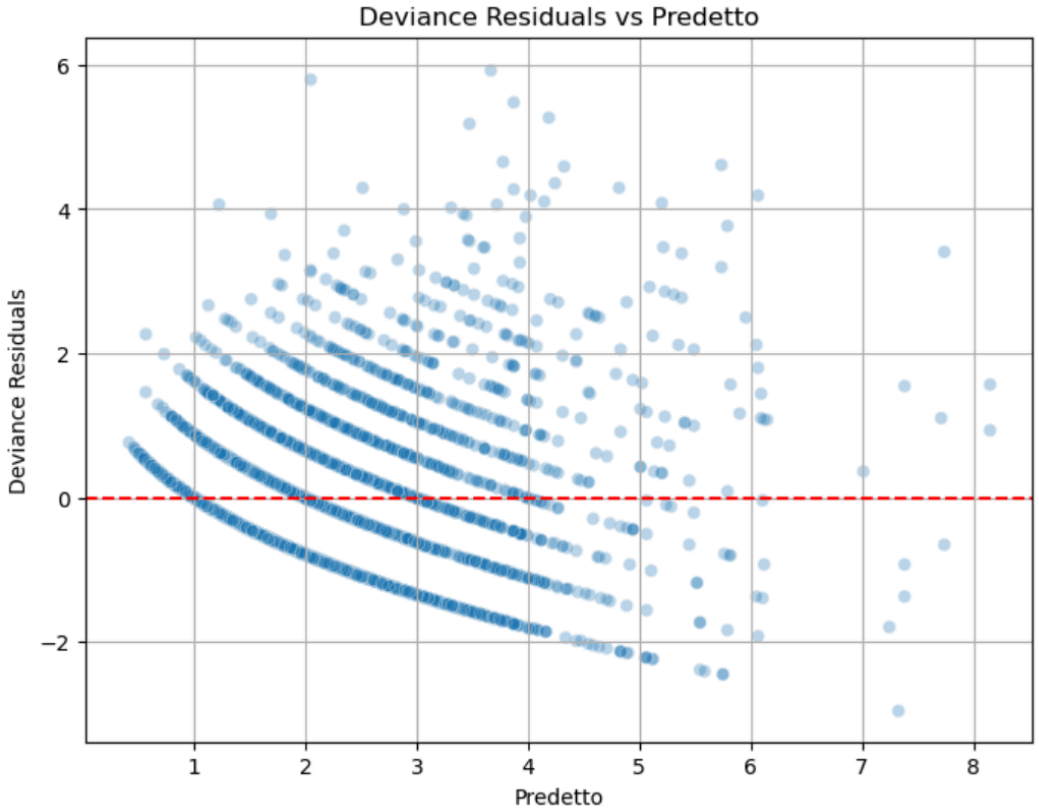
At first glance, most points lie in a compact zone between 1 and 5 for predicted sales, and 1 to about 10 for observed sales. This matches the typical sales structure in fashion luxury, where few SKUs exceed high sales thresholds, while the vast majority fall into low to medium volumes.

The model seems to focus well on the core of the distribution, where most sales occur.
Some vertical dispersions are visible, for example, for predicted values around 3 or 4, where observed values reach 15 or 20. These cases are natural outliers, i.e., SKUs that performed exceptionally well compared to the average.

The model's inability to fully capture these peaks is expected, as Poisson regression tends to estimate the mean rather than extreme deviations.

Overall, the graph suggests good coherence between predictions and reality. The resulting triangular shape (wider at the base, narrower at the top) is typical of well-calibrated Poisson models on discrete, skewed data.
The presence of significant concentration in the lower observed and predicted values confirms that the model fits well in the core business zone, while for exceptional values there remains room for future improvements, such as through more flexible models or additional predictive covariates.

Deviance Residuals vs Predicted Plot



This graph shows the relationship between the predicted values from the Poisson regression model (X-axis) and the corresponding deviance residuals (Y-axis). Deviance residuals measure how much each data point deviates from the model's prediction: the farther a point is from the red horizontal line at zero, the greater its deviation.

The pattern shows a fan-shaped distribution upwards, with residual values increasing as the predicted value increases. This behavior is quite common in Poisson models, especially when working with discrete data with a strong concentration in low values, as in fashion luxury sales.

The horizontal layering effect is due to the discrete nature of the response variable, sold_qty, and the fact that many real values are identical (e.g., many SKUs sold 1 or 2 units), even though their predictions vary.

The red dashed line represents perfect model adherence (residual = 0). Ideally, in a perfect model, the points would be symmetrically and randomly

distributed around this line. In this case, the residuals do not show a strongly irregular structure, but there is a slight upward trend with the prediction.
This may suggest the model slightly overestimates or underestimates certain product segments

4.2.3 Rationale for Outlier Analysis

After analyzing the entire dataset with a Poisson regression extended to all observations, it was decided to further investigate the model's behavior in the presence of high-selling SKUs. These cases, which represent the positive outliers in the distribution of the sold_qty variable, are few in absolute terms but highly relevant in terms of profit, positioning, and commercial strategy.

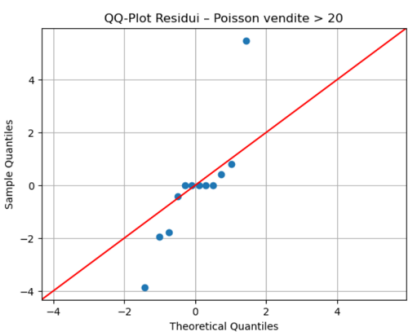
In the fashion luxury sector, the ability to intercept and forecast these outliers is central: they often represent best-sellers or driving SKUs, or those pushed by strong advertising investment, whose correct distribution can decisively influence the entire collection.

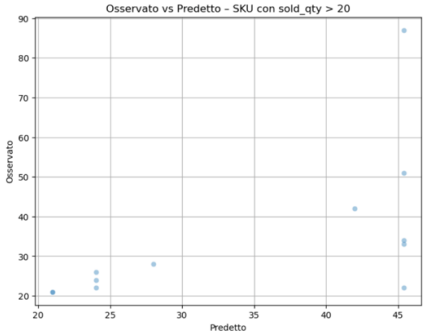
The objective of this section is therefore to verify whether the Poisson model, even in the absence of complex numerical variables, is able to capture the predictive signals associated with high-selling products.

In terms of overall performance, the model shows a pseudo R-squared (Cox & Snell) of 0.9685, a particularly high value indicating strong explanatory power, although on a limited dataset. The log-likelihood (-58.048) and deviance (52.773) also appear consistent with a good fit, considering the small number of observations. The Pearson Chi^2 is 57.5, while the residual degrees of freedom are 5: an indicator of low dispersion, suggesting good adherence between observed and predicted data, without excessive problems of overdispersion.

The model offers a consistent and predictively useful framework to assess the expected behavior of high-potential SKUs. The coefficients, largely significant,

provide valuable information for the merchandising team and regional planners, who could, based on this evidence, define in advance a more suitable set of products for each market.





The first graph , the QQ-Plot of residuals, shows that the deviance residuals are fairly well distributed along the theoretical normal quantile line, suggesting good model stability even when handling complex or extreme cases. The presence of a few marginal points deviating from the line is physiological, especially considering the small sample size and the extreme variability of top SKU behavior, but it does not undermine the overall structure of the model.

The second graph, which compares observed and predicted value, confirms that, despite slight deviations, the model is able to reliably estimate the order of magnitude of sales for each of the analyzed SKUs. There are no extreme outliers or gross errors, and most importantly, a consistent relationship is observed between predictions and actual sales, reinforcing the idea that the model is already capable of supporting strategic decisions for core products, at least in a comparative logic between SKUs.

These two graphs, analyzed together, demonstrate that the model not only works as a whole but also maintains solidity in the most important and difficult cases – those where business decisions have the greatest economic and operational impact. In summary, it can be said that the modeling base is robust, interpretable, and already useful to support the buying process, while still leaving room for future improvements.

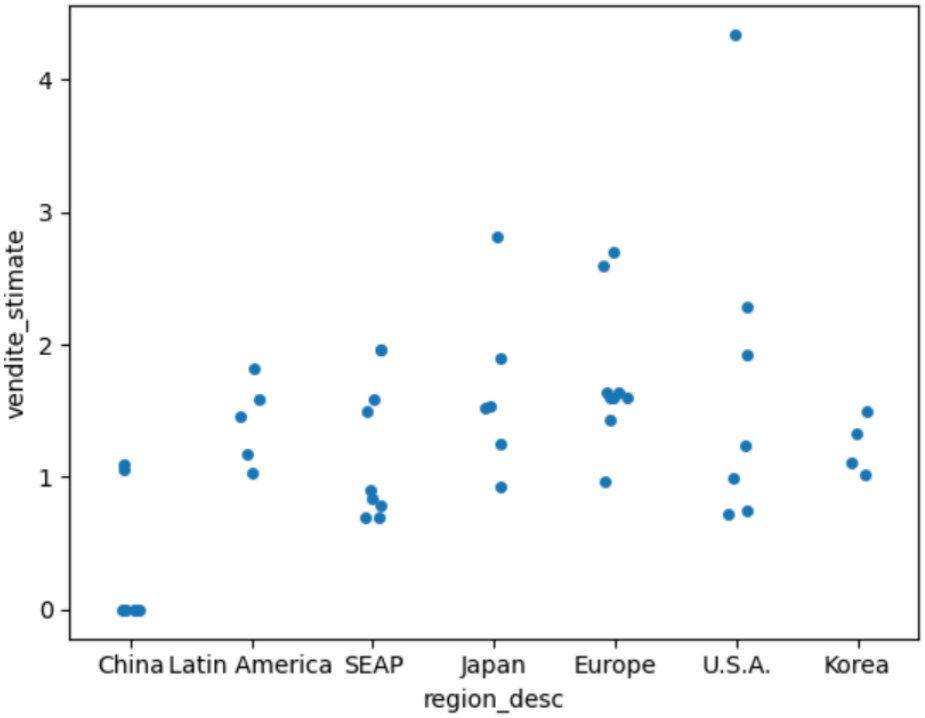
4.3 Simulation on Fictitious SKUs: A Predictive Model Validation Exercise

After thoroughly exploring the behavior of the Poisson model on real data, this file takes on a crucial role in completing the validation process. In this section, the focus shifts to a simulation designed to observe whether the algorithm, once fed with new combinations of product and store characteristics, is actually able to generate predictions consistent with what was learned during training. This is therefore not a mere ex-post forecast based on known data, but a true qualitative verification of the model's behavior in a simulated and controlled context, where variables were specifically constructed to test its robustness.

The idea behind this experiment is simple: a small dataset is built, composed of "fictitious" SKUs, not present in the original database, but constructed according to realistic logics consistent with the typical configurations of Ferragamo's collections. Each row represents a hypothetical handbag, characterized by plausible values for all the categorical variables used in the model (such as color, product positioning, size, line, or store type), but with the sold_qty variable to be predicted. At this point, the previously trained Poisson model on real data is applied, and an estimate of the predicted quantity sold (sold_qty_predicted) is generated.

There is a very careful selection in the composition of the simulated combinations. For example, some SKUs were deliberately defined with high-turnover characteristics observed in historical data, such as a macro_line_grouped associated with successful lines, or a price_range in the medium-high tier, within regions that have historically shown greater purchasing propensity. In other cases, "critical" or weak configurations in terms of performance are simulated, such as a culturally less successful color in a given country or a size-line combination that doesn't align with the preferences of that particular clientele.

This alternation between high and low potential SKUs allows the model to be stress-tested and its predictive sensitivity verified: the point is not only to observe how high or low the predicted values are, but to understand whether they reflect the logic learned during the training phase – in other words, whether the score assigned by the model to various configurations is consistent with the patterns observed in the historical dataset.



From the results, a very encouraging behavior emerges. For example, SKUs placed in regions like Europe or U.S.A., associated with medium sizes, neutral colors, and ongoing lines, tend to have significantly higher sold_qty predictions, confirming the model's ability to recognize, even in a hypothetical scenario, the favorable conditions for selling. Conversely, less "coherent" combinations, such as large handbags in SEAP or flashy colors placed in classically positioned store clusters, generate more cautious predictions, often below 3 units estimated, indicating that the algorithm has learned a sensible predictive structure, even though it has never "seen" these SKUs during training.

Among the results, the most significant value is that "Ferragamo HUG" was rightly predicted as a possible best seller in almost all regions, confirming the real value this handbag has generated.

A methodologically important point should be highlighted: the Poisson model, as built, is highly dependent on the categorical relationships between variables. In other words, the entire predictive process is based on estimating the expected sales intensity for each combination of levels present in the model's categorical variables. Since all the selected variables for the final model were encoded as factors (C(...)), the prediction behaves like an additive overlay of the average effects observed for each level. This is consistent with the approach of generalized Poisson regression, but at the same time introduces some limitations, as will be noted in the chapter conclusions.

An interesting aspect of this simulation phase concerns the visual comparison between SKUs: a direct comparison is made between SKUs that differ by only one characteristic (for example, same store and same line but different colors). In these cases, the model is observed to penalize, even substantially, less "appetizing" color variants for that specific context, assigning them lower predicted sales. This is a fundamental dynamic for real-world business applications, where assortment choices are based precisely on selecting the most promising combinations within similar product families.
Naturally, there are limitations and possible improvement directions. As noted in previous analyses, the model's entirely categorical nature requires that each combination be present in the training set to be "recognized" and properly valued. In a real-world scenario, this could limit generalization capability to truly new SKUs or unseen combinations, thus pushing towards more flexible models such as the GBM regressor.
However, in the context of this simulation, the model's performance was consistent with the predefined goals: to demonstrate the validity of the predictive approach adopted and its ability to replicate decision-making patterns in a controlled environment. Even with a simple structure, the model manages to "reason" in terms of expected performance, returning outputs that respect the logic observed in reality and that could, with further refinement, become a valid support for the SKU selection phase within the merchandising process.

In the next chapters, it will be interesting to observe how this approach performs on a new data base, namely the one referring to actual handbag performance, to verify whether the same predictive logic will be able to provide a faithful representation of the demand observed in the market.

CHAPTER 5 – Final Application on Real Ferragamo Data: Poisson Prediction and Binary Classification of Handbag SKUs

5.1 Context: Towards a Data-Driven Strategy in SKU Selection

region_desc	dimension_group_desc	macro_line_grouped	colour_group	price_range	macro_merc_typology_desc	sku_code	macro_colour_group
U.S.A.	Medium	THE STUDIO	WHITE	HIGH	Tote	786555	NEUTRAL
U.S.A.	Medium	THE STUDIO	BROWN	HIGH	Tote	786556	BROWN
U.S.A.	Large	THE STUDIO	BLACK	HIGH	Tote	786557	NEUTRAL
U.S.A.	Large	THE STUDIO	MULTICOLOUR	TOP	Tote	786558	NaN
U.S.A.	Large	FERRAGAMO HUG	DARK BROWN	HIGH	Top Handle	786559	BROWN
U.S.A.	Medium	FERRAGAMO HUG	MEDIUM BROWN	HIGH	Top Handle	786560	BROWN
U.S.A.	Mini	FERRAGAMO HUG	LIGHT YELLOW	CORE HIGH	Minibag	786561	YELLOW_GOLD
U.S.A.	Medium	FERRAGAMO HUG	BLACK	HIGH	Shoulder Bag	786562	NEUTRAL
U.S.A.	Large	FERRAGAMO HUG	DARK BROWN	HIGH	Top Handle	786563	BROWN
U.S.A.	Mini	FERRAGAMO HUG	LIGHT YELLOW	CORE MEDIUM	Minibag	786564	YELLOW_GOLD

This section represents the most significant and ambitious phase of the entire thesis, as it marks the transition from experimental analyses and simulations to a test on real data provided by the company Salvatore Ferragamo. The underlying objective remains consistent with what was declared in the previous chapters: to develop a predictive model capable of effectively supporting decisions related to the selection of SKUs in the handbag category, optimizing the decision-making process that precedes seasonal buying and significantly reducing time-to-market.

In the fashion luxury segment, characterized by strong geographical, seasonal, and cultural fragmentation, the ability to anticipate which products are more likely to sell is a crucial element to ensure logistical efficiency and consistency of brand image. However, the major challenge is not to estimate exactly the number of units sold (continuous target), but rather to identify those SKUs that, within the proposed assortment, possess characteristics that make them strategically suitable to be purchased by buyers for the various regions. The emphasis, therefore, shifts from mere quantitative forecasting to qualitative prioritization: not “how many pieces will go into a store or will be sold,” but “which items are likely to perform well in the target markets.”

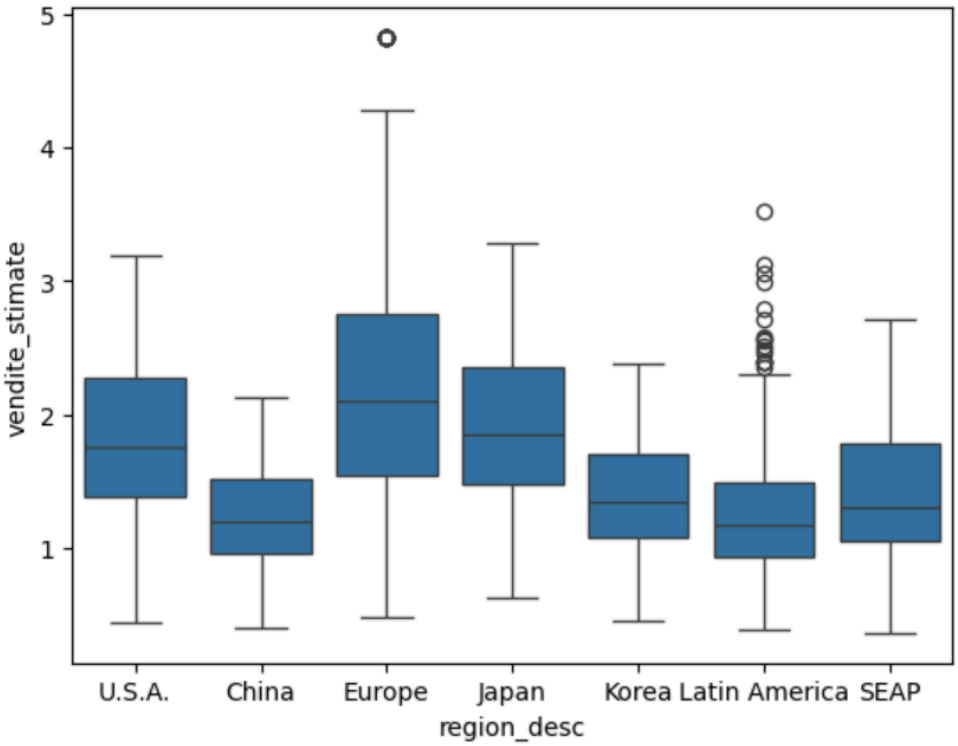
With this goal, two parallel approaches were developed:
A Poisson regression model (GLM), aimed at estimating the quantity sold for each SKU based solely on categorical variables.
A binary classification model, aimed at distinguishing between selling and non-selling SKUs (binary target: 1 if sales > 0, 0 otherwise).

Both models were tested on a selected subset of the main database, related to handbag SKUs for each store and region, and focus only on the information available before product distribution: such as price range, product line, color, size, region, and store.

Explanatory Analysis of the Results, Poisson Model on Real Ferragamo SKUs (Handbags)

After applying the Poisson Generalized Linear Model (GLM) to the real dataset relating to the handbags category SKUs, distributed across Ferragamo's international stores, we finally reach the most important phase of the entire thesis: the strategic interpretation of the results obtained, and the assessment of their potential impact on the company's decision-making process. This is not about evaluating the numeric value of the estimated sales, but rather about understanding what predictive signals emerge, how these differ by geographical area, and above all how such an algorithm can concretely assist buyers and planners in improving the global assortment, reducing time-to-market and preventing unsold inventory.

Reading the Results



One of the main observations is that each region tends to show a peculiar demand behavior, meaning that the propensity to sell the same SKU varies significantly from one geographical area to another. This variability is an expected result, and represents one of the main reasons why variables such as region_desc, country, store_code, and other categorical dimensions were introduced into the model.

For example, some SKUs have a high probability of sale in Asian stores (e.g. China and Japan), but weak performance in SEAP stores. In other cases, SKUs well received in Europe or the U.S.A. do not achieve the same success in Korea. This clearly reflects the effect of cultural preferences, inverse seasonality, and preferred sizes and product lines in each market.

This evidence demonstrates two key concepts:
Uniform distribution is unrealistic: thinking that the same assortment can work across all regions is an outdated assumption.
Personalization is essential: each region requires dedicated study, with SKUs selected based on their predictive probability of selling.
But this is precisely the problem that arises when talking about GCA.

Identification of High-Potential Cross-Regional SKUs

Despite the strong regional specificity, another important point emerged in the final part of the analysis: the presence of some common SKUs that show good predictive performance across multiple regions. These SKUs represent a strategic asset for the company: they are “global” products, capable of performing well regardless of geographical location, and are therefore ideal candidates to be included in the Global Core Assortment (GCA).

	region_desc	store_code	sku_code	prob_vendita
10	China	A01N	786565	0.999970
29	China	A01N	786584	0.999849
27	China	A01N	786582	0.999829
9	China	A01N	786564	0.999814
15	China	A01N	786570	0.999734

	store_code	sku_code	vendite_stimate
15	A01N	786570	2.127900
34	A01N	786589	1.886862
32	A01N	786587	1.845652
31	A01N	786586	1.810551
16	A01N	786571	1.685313
54	A516	786571	2.636579
73	A516	786590	2.337877
71	A516	786588	2.286789
70	A516	786587	2.243379
55	A516	786572	2.088254

This is perhaps the most practical utility of the applied Poisson model: it not only identifies which SKUs are better suited for a specific store, but also helps to select the more solid, transversal, and universal ones, which can guarantee positive results in multiple markets. In other words, the model can support the company not only in local personalization, but also in global strategic standardization.

This point is crucial for Ferragamo, which for years has been trying to balance international brand coherence with responsiveness to regional preferences. All of this is then confirmed by the classification model, which supports the identification of SKUs with higher sales probability.

The model, for example, recognizes a HUG in mid brown as a best seller in few regions.

Bianca Balti with Ferragamo hug in Sanremo 2024

new Ferragamo HUG in midbrown

Possible bestseller.

Impact on Governance: Reducing Time-to-Market

One of the declared objectives of this thesis, since the introduction, was to reduce time-to-market, that is, to shorten the time between product design and its actual arrival in store. This issue is closely linked to the assortment decision process, which currently requires numerous meetings between planners and buyers, manual analysis of Excel files, and long negotiations between regions.

The use of the GLM model allows:

•To anticipate SKU evaluation even before production, based on their physical and categorical characteristics,

•To provide planners with an objective, fast, and visual support, thanks to estimated sales probabilities and performances,

•To streamline the comparison process with regions, facilitating the selection of the most promising SKUs.

The result is faster decision-making, better alignment between style and demand, and a significant reduction in unsold risk, as already discussed in Chapter 2.

From Model to Business: Concrete Examples of Impact

To better understand the strategic value of this analysis, it is useful to hypothesize some concrete situations in which the model could be used within the company:

•A planner must choose between two similar bag models, but with different color variants: the predictive model shows that the “bordeaux” variant has a higher probability of sale in Europe and Japan, while the “white” version performs well only in one region. The planner may suggest the first one as part of the GCA.

•A region insists on purchasing 10 SKUs not validated by historical data: the Poisson model, based on structural variables, provides a very low estimate for all 10. The planner can use this data to propose better alternatives.
•During the new season planning, the style department proposes 20 new models: the merchandising team can pass them through the model to estimate their performance and discard early those with critical estimates.

Coherence with Business Strategy and Future Upgrades

Finally, while recognizing that the current model has limitations (e.g., lack of behavioral variables such as sell_through or more granular transactional data), it is important to underline that the objective of this thesis was to demonstrate the validity of the predictive approach, not to build a definitive system.

Future upgrades may include:
•Introduction of temporal variables (e.g., week of sale),
•Use of behavioral features (e.g., returns, refunds, discounting, advertising), use of external clouds to assess the market,
•Integration with probabilistic classifiers or recommendation systems.

With these steps, the model could offer real value: it improves governance, provides visual and measurable insights, and helps Ferragamo make faster and more informed decisions.

Possible Gaps and Future Evolutions of the Model

Although the results obtained may confirm the validity of the approach adopted, it is essential to highlight that the system currently developed represents a first step in a broader path of organizational and cultural transformation. This chapter does not aim to highlight limitations in a negative sense, but rather to bring attention to areas of strategic improvement, which could significantly enhance the model's accuracy, scalability, and operational impact in the medium-long term.

GAP 1 – Database Structure: from disordered historical data to a model-oriented design

One of the main limitations encountered during the study concerns the heterogeneous and poorly standardized nature of the available data. Although the real datasets provided by Ferragamo are rich in terms of history and volume, the lack of coherent structure and homogeneous nomenclature represented a major obstacle during the pre-processing phase and especially in the stability of the model.

Many categorical fields (e.g. macro_line_grouped, macro_merc_typology_desc, macro_colour_group) present different labels for similar concepts, or clusterings and naming conventions that vary over time. This reduces the model's ability to learn consolidated patterns and makes training across multiple seasons more difficult.

Proposed evolution:

- Design a fixed, well-thought-out, and machine learning-oriented data structure. Every new collection should adopt stable codings, coherent formats, and a controlled taxonomy.
- Introduce predefined micro-clusters (e.g. shape, color palette, size) that serve as synthetic and stable features over time.

This would allow for greater model generalizability, a reduction in cleaning time, and, most importantly, the ability to use more sophisticated models based on transfer learning between seasons.

For predictive models to be effective over time, they must be trained on data that do not change structure with every campaign.

GAP 2 – Lack of an internal structure dedicated to predictive analysis and AI-driven strategy

The second gap concerns the organizational structure of the fashion luxury sector, which today, in companies like Ferragamo, still does not foresee an internal division dedicated to the research and development of predictive models or the strategic use of artificial intelligence.

Unlike more data-driven industries such as beauty (P&G) or automotive (Volkswagen, Tesla), in fashion, the "data science" function is often still embedded within business or demand planning departments, with heterogeneous skills not always focused on machine learning and AI. The analysis is often retrospective, while a strongly prospective, simulation-based, and experimental approach is missing.

Proposed evolution:
•Creation of a cross-functional "Data & AI" team, with hybrid profiles among data scientists, merchandisers, market analysts, and model developers.
•Introduction of internal labs or pilot projects, in collaboration with universities or research centers (e.g. LUISS Data Lab), to develop agile prototypes on selected datasets.
•Integration of interactive dashboards and what-if simulation systems to support the work of planners and buyers.

In luxury, the culture of data is still young: investments may be needed, not only in software, but also in human capital and strategic vision. As already demonstrated in the first chapter.

GAP 3 – Model evolution towards more performant algorithms: a first test with Gradient Boosting

The Poisson model (GLM) has shown good descriptive and interpretative capabilities, but it still represents a linear and parametrically simple model. In the final part of the project, an internal test was conducted on a subset of real data, where a Gradient Boosting Machine (GBM) model was implemented for sales forecasting.

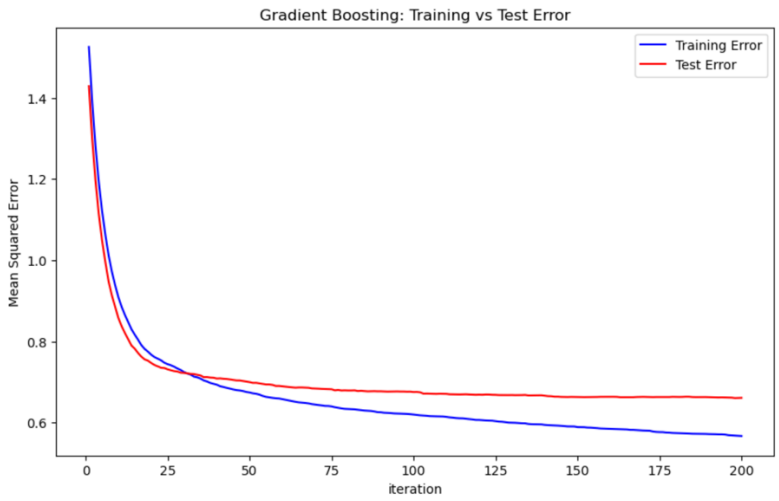
Compared to the GLM, this model presents three main advantages:

•Greater ability to capture non-linearity and complex interactions between variables, which is very useful in a dataset with strong categorical components and implicit relationships.

•Lower risk of multicollinearity or parametric overfitting, since the algorithm builds the prediction function through successive approximations.

•Possibility to calculate the relative importance of variables in a more robust way.

```
R²: 0.5829640227871912
MAE: 0.31058402589319045
RMSE: 0.8129435689945892
```



This how GBM works on the same database of Poisson GLM, good result to be examined in more detail

Proposed evolution:

•Experiment with the use of alternative machine learning models, such as LightGBM or XGBoost, in a production-like environment with interactive dashboards.

CONCLUSIONS

Conclusion: a path to consolidate, not to abandon

The adoption of predictive models in fashion luxury does not represent a passing trend, but a strategic necessity to govern the complexity of a globalized, dynamic sector, strongly demand-driven, yet planned months in advance.

It is true that in recent years many investments have been made in data-driven technologies in the fashion sector as well, often with inconsistent results. Today, however, the maturation of data analysis and artificial intelligence tools has opened up new concrete perspectives.

Fashion will never be an exact science: a single unexpected event, such as a celebrity wearing a specific bag, is enough to alter sales trends in just a few days, nullifying every forecast.

The Italian case, and in particular that of Ferragamo, strongly reflects these dynamics. The sector is going through a critical phase that, in many cases, translates into a direct impact on employment and on the resilience of the manufacturing fabric, with consequent risk for the artisanal know-how that has made Italian fashion famous worldwide.

A more innovative planning process, supported by empirical studies and predictive models, could help companies significantly reduce the risk of strategic mistakes.

It is time to start looking at the term “innovation” not only from a stylistic point of view.

LITERATURE

(s.d.).

Afonso, P. N. (2008). *The influence of time-to-market and target costing in the new product development success.* (Vol. 115(2)). International Journal of Production Economics.

Agency, I. E. (2024). *Trends in electric cars. In Global EV Outlook 2024.* Tratto da https://www.iea.org/reports/global-ev-outlook-2024/trends-in-electric-cars

al., S. e. (2020). *Comparison of statistical and machine learning methods for daily SKU demand forecasting.*

Bacconi, A. D. (2022). *Machine learning and artificial intelligence use in marketing: a general taxonomy.* Italian Journal of Marketing .

Bof-McKinsey . (2025). *The State of Fashion 2025.* Bof-McKinsey .

Camur, M. C. (2024). *Enhancing supply chain resilience: A machine learning approach for predicting product availability dates under disruption. .* Expert Systems with Applications.

Chen-Yu, J. H. (2020). *Consumer characteristics as predictors of purchase intentions and willingness to pay a premium for men's mass-customized apparel.* Journal of Global Fashion Marketing.

Company., M. &. (2023). *McKinsey & Company.* Tratto da The State of Fashion Technology. Business of Fashion & McKinsey & Company Report: https://www.mckinsey.com/

De Mauro, A. S. (2022). *Machine learning and artificial intelligence use in marketing: A general taxonomy.*

Kalhor, M. &. (2022). *Smart Fashion: A Review of AI Applications in the Fashion & Apparel Industry.*

KNIME. (2025). *KNIME*. Tratto da https://www.knime.com/success-story/how-pg-uses-real-time-data-supply-chain-resiliency

KPMG. (2024). *The Market of luxury goods.*

Kumar, A. S. (2024). *A novel approach for prediction of consumer buying behavior of luxury fashion goods using machine learning algorithms. International Journal of Intelligent Systems and Applications in Engineering.*

McDowell, C. (. (2024). Luxury Goods Worldwide Market Study, Spring–Summer.

McDowell, C. (2000). *Fashion Today. Phaidon.*

Ogundipe, D. O. (2024). *AI and product management: A theoretical overview from idea to market.* International Journal of Management & Entrepreneurship Research.

SAM RANSBOTHAM, F. C. (2021). *The Cultural Benefits of Artificial Intelligence in the Enterprise.* MIT sloan.

Shoaib, M. (2024, November 13). *Vogue Business*. Tratto da Vogue: https://www.voguebusiness.com/story/consumers/luxurys-growth-stutters-as-50-million-consumers-pull-back-on-spending

Statista. (2024). AI in Fashion - Global Market Forecast 2022–2027.

Statista. (2024). BEAUTY TECH: MARKET DATA & ANALYSIS.

Stumpf, R. (2024). *Insideevs*. Tratto da Insideevs: https://insideevs.com/news/748876/vw-id7-canceled-north-america/

W. H. Thejani Madhuhansi, L. K.-M. (2025). *Consumer Disposal Behaviour of Durable and Semi-Durable Products: A Systematic Literature Review and Future Research Agenda.* International Journal of Consumer Studies.

Y., Z. (2023). *Dalla tradizione alla tecnologia: esplorare l'evoluzione del commercio al dettaglio di lusso e la frontiera digitale.*

SOURCES

Ogundipe, D. O., Babatunde, S. O., & Abaku, E. A. (2024). AI and product management: A theoretical overview from

idea to market. International Journal of Management & Entrepreneurship Research, 6(3), 950-960

Usman, F. O., Eyo-Udo, N. L., Etukudoh, E. A., Odonkor, B., Ibeh, C. V., & Adegbola, A. (2024). A critical review of

ai-driven strategies for entrepreneurial success. International Journal of Management & Entrepreneurship

Forbes. (2023, February 21). Artificial intelligence in fashion. Retrieved fromhttps://www.forbes.com/councils/theyec/2023/02/21/artificial-intelligence-in-fashion/

International Journal of Supply Chain Management, How AI Adapts to Market Disruptions and Consumer Shifts (2023)

InsightAce Report – AI in Beauty and Cosmetics Market (2025)

Revieve. (2021, November 1). Percentage of Gen Z consumers who trust AI advisors for personalized skincare recommendations in North America in 2021 [Graph]. Statista. https://www.statista.com/statistics/1289772/gen-z-s-trust-in-ai-beauty-advisors-in-north-america/

Mohammed, I. A., & Mandal, J. (2022). Forecasting Accuracy through Machine Learning in Supply Chain Management. International Journal of Supply Chain Management, 7(2), 60-77.

Kumari, D., & Bhat, S. (2021). Application of artificial intelligence technology in tesla-a case study. International

Journal of Applied Engineering and Management Letters (IJAEML), 5(2), 205-218.

Ma, H., He, B. Y., Kaljevic, T., & Ma, J. (2024). A two-sided model for EV market dynamics and policy implications.

Bain & Company. (2024). Luxury Goods Worldwide Market Study, Spring–Summer 2024. Bain & Company.

Ferragamo S.p.A. (2024). Relazione finanziaria annuale consolidata 2023. Firenze: Salvatore Ferragamo Group.

Ren, S., Chan, HL. & Siqin, T. Demand forecasting in retail operations for fashionable products: methods, practices, and real case study.

Spiliotis et al. (2020) – Comparison of statistical and machine learning methods for daily SKU demand forecasting

Haque et al. (2023) – Retail Demand Forecasting: A Comparative Study for Multivariate Time Series

Gonçalves, J. N. C., Carvalho, M. S., & Cortez, P. (2023). Operations research models and methods for safety stock determination: A review. Computers & Industrial Engineering Research, 6(1), 200-215.

Swink, M., Melnyk, S. A., Hartley, J. L., & Cooper, M. B. (n.d.). Managing operations across the supply chain (4th ed.). McGraw-Hill Education.

APPENDIX

PYTHON CODE

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
pd.set_option ('display.max_columns', 500)
```

In [2]:

```
df =
pd.read_excel("/Users/stefanolandolfi/Desktop/db_usa_handbags_22-
23-24_price_range.xlsx", sheet_name="handbags_usa")
df.head()
```

Out[2]:

controllo se i valori di received sono tutti univoci

In [3]:

```
df['RECEIVED'].unique()
```

Out[3]:

```
array([  2.,   1.,   4.,   3.,   5.,   0.,  -1.,   6.,   9.,   7.,
8.,
        10.,  12.,  14.,  11.,  13.,  15.,  19.,  17.,  20.,  24.,
33.,
        16.,  18.,  30.,  28.,  27.,  39.,  54.,  25.,  42.,  21.,
45.,
        50.,  48.,  31.,  23.,  43.,  67.,  52.,  36.,  37.,  32.,
58.,
       119.,  57.,  22.,  26.,  35.,  97.,  41.,  46.,  63.,  74.,
55.,
        56.,  68.,  89.,  51.,  29., 195.,  77., 186.,  38.,  71.,
70.,
       219.,  64., 114.,  69.,  60., 108.,  96.,  34., 291.,  40.,
78.,
       221.,  72., 179.,  49.,  47.,  59.,  62.,  93.,  76., 177.,
nan])
```

In [4]:

```
print(df['RECEIVED'].dtype)
print(df['RECEIVED_LY'].dtype)
print(df['RECEIVED_LLY'].dtype)
float64
int64
int64
```

come si può notare received non è intero, si deve cambiare però prima di fare questo si devono rimuovere i valori NaN

In [5]:

```
df=df.loc[(df['Stock Program Flag Market'].isna())&(df['Collection
Stock Program Flag'].isna())].copy()
```

Filtra il DataFrame df e mantiene solo le righe in cui entrambe queste condizioni sono vere

In [6]:

```
df.shape
```

Out[6]:

```
(68871, 52)
```

In [7]:

```
df.columns = df.columns.str.strip().str.lower().str.replace(' ',
'_')
```

In [8]:

```
print(df.columns.tolist())
['region_desc', 'country', 'channel_desc', 'store_code',
'store_code_desc', 'entity_typology', 'sku_birth_collection',
'prod_cat_gender_desc', 'rotb_macro_category_desc',
'production_category_code', 'prod_category_code_+_desc',
'macro_merc_typology_desc', 'merchandise_typology_desc',
'macro_line_desc', 'line_desc', 'model_code', 'model_name',
'color_desc', 'material_macro_group_desc', 'material_group_desc',
'dimension_group_desc', 'construction_group', 'occasion_desc',
'sku_code', 'sku_code_+_desc', 'special_production_desc',
'brand_desc', 'creation', 'gca', 'stock_program_flag_market',
'collection_stock_program_flag', 'sku_type', 'price_list_eur',
'net_sold_qty_by_coll_rtl_ltd', 'net_sold_qty_by_coll_rtl_ltd_ly',
'net_sold_qty_by_coll_rtl_ltd_lly',
'net_rev_excl_vat_by_coll_rtl_ltd',
'net_rev_excl_vat_by_coll_rtl_ltd__ly',
'net_rev_excl_vat_by_coll_rtl_ltd__lly', 'microcoll', 'received',
'received_ly', 'received_lly', 'key', 'store_code',
'regioncountry', 'sell_tr', 'sell_tr_ly', 'sell_tr_lly',
'cluster', 'price_range', 'colour_group']
```

In [9]:

```
df['received'] = df['received'].fillna(0)
```

In [10]:

```
df['sell_tr'] = df['sell_tr'].fillna(0)
```

In [11]:

```
df.isna().sum(axis=0)
```

Out[11]:

```
region_desc                          0
country                              0
channel_desc                         0
store_code                           0
store_code_desc                      0
entity_typology                      0
sku_birth_collection             25163
prod_cat_gender_desc                 0
rotb_macro_category_desc             0
```

```
production_category_code                    0
prod_category_code_+_desc                   0
macro_merc_typology_desc                    0
merchandise_typology_desc                   0
macro_line_desc                             0
line_desc                                   0
model_code                                  0
model_name                                  0
color_desc                                  0
material_macro_group_desc                   0
material_group_desc                         0
dimension_group_desc                        0
construction_group                      68871
occasion_desc                               0
sku_code                                    0
sku_code_+_desc                             0
special_production_desc                     0
brand_desc                              25163
creation                                25163
gca                                     68871
stock_program_flag_market               68871
collection_stock_program_flag           68871
sku_type                                    0
price_list_eur                              0
net_sold_qty_by_coll_rtl_ltd                0
net_sold_qty_by_coll_rtl_ltd_ly             0
net_sold_qty_by_coll_rtl_ltd_lly            0
net_rev_excl_vat_by_coll_rtl_ltd        43708
net_rev_excl_vat_by_coll_rtl_ltd__ly    25163
net_rev_excl_vat_by_coll_rtl_ltd__lly   25163
microcoll                               43708
received                                    0
received_ly                                 0
received_lly                                0
key                                         0
store_code                               7582
regioncountry                            7582
sell_tr                                     0
sell_tr_ly                                  0
sell_tr_lly                                 0
cluster                                  7805
price_range                                 0
colour_group                                0
dtype: int64
```

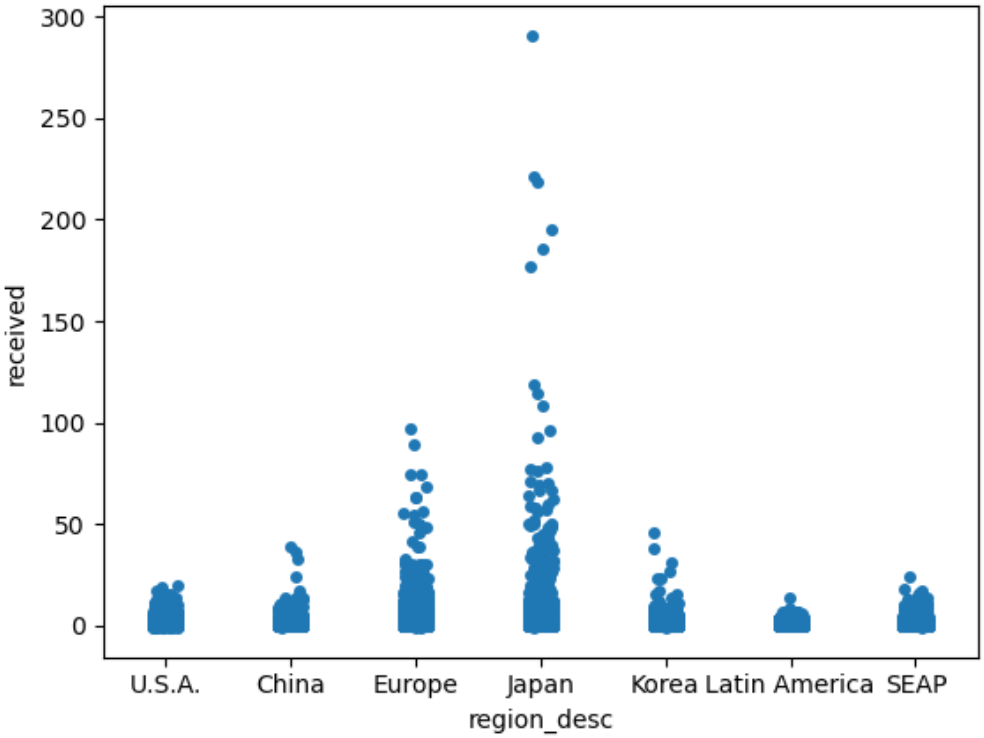
negli ultimi 3 passaggi ho trasformato tutti i received e i selltr NaN in 0

In [15]:

```
sns.stripplot(data=df, x='region_desc', y='received')
```

Out[15]:

```
<Axes: xlabel='region_desc', ylabel='received'>
```



In [16]:

```
# Clip per sell-through: tra 0 e 1
selltr_cols = ['sell_tr', 'sell_tr_ly', 'sell_tr_lly']
df[selltr_cols] = df[selltr_cols].clip(lower=0, upper=1)

# Clip per received e net_sold_qty: solo valori >= 0
other_cols = [
    'received', 'received_ly', 'received_lly',
    'net_sold_qty_by_coll_rtl_ltd',
    'net_sold_qty_by_coll_rtl_ltd_ly',
    'net_sold_qty_by_coll_rtl_ltd_lly'
]
df[other_cols] = df[other_cols].clip(lower=0)
```

ho settato i valori di sell tr tra 0 e 1 poiché è una percentuale, poi i received e le vendite >=0 perché nel db sono compresi anche i resi ma non è giusto considerarli

In [17]:

```
columns_to_check = [
 'net_sold_qty_by_coll_rtl_ltd',
 'net_sold_qty_by_coll_rtl_ltd_ly',
 'net_sold_qty_by_coll_rtl_ltd_lly',
 'received',
 'received_ly',
 'received_lly',
 'sell_tr',
 'sell_tr_ly',
 'sell_tr_lly']
df[columns_to_check].describe()
```

Out[17]:

In [18]:

```
df['received'] = df['received'].astype(int)
```

In [19]:

```
df.head()
```

Out[19]:

ho trasformato il valore di received in intero

In [20]:

```
df['sell_tr'].isna().sum()
```

Out[20]:

```
np.int64(0)
```

In [21]:

```
columns_to_check = [
 'net_sold_qty_by_coll_rtl_ltd',
 'net_sold_qty_by_coll_rtl_ltd_ly',
 'net_sold_qty_by_coll_rtl_ltd_lly',
 'received',
 'received_ly',
 'received_lly',
 'sell_tr',
 'sell_tr_ly',
 'sell_tr_lly']
df[columns_to_check].describe()
```

Out[21]:

	net_sold_qty_by_coll_rtl_ltd	net_sold_qty_by_coll_rtl_ltd_ly	net_sold_qty_by_coll_rtl_ltd_lly	received	received_ly	received_lly	sell_tr	sell_tr_ly	sell_tr_lly
count	68871.000000	68871.000000	68871.000000	68871.000000	68871.000000	68871.000000	68871.000000	68871.000000	68871.000000
mean	0.214648	0.163029	0.155595	0.951445	1.198313	1.113865	0.072186	0.049841	0.044451
std	0.971706	0.628812	0.573185	3.266187	2.119110	2.594571	0.223357	0.186520	0.169871
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	2.000000	2.000000	2.000000	0.000000	0.000000	0.000000

	net_sold_qty_by_coll_rtl_ltd	net_sold_qty_by_coll_rtl_ltd_ly	net_sold_qty_by_coll_rtl_ltd_lly	received	received_ly	received_lly	sell_tr	sell_tr_ly	sell_tr_lly
max	87.000000	18.000000	13.000000	291.000000	151.000000	278.000000	1.000000	1.000000	1.000000

In [22]:

```
print(df.columns.tolist())
['region_desc', 'country', 'channel_desc', 'store_code',
'store_code_desc', 'entity_typology', 'sku_birth_collection',
'prod_cat_gender_desc', 'rotb_macro_category_desc',
'production_category_code', 'prod_category_code_+_desc',
'macro_merc_typology_desc', 'merchandise_typology_desc',
'macro_line_desc', 'line_desc', 'model_code', 'model_name',
'color_desc', 'material_macro_group_desc', 'material_group_desc',
'dimension_group_desc', 'construction_group', 'occasion_desc',
'sku_code', 'sku_code_+_desc', 'special_production_desc',
'brand_desc', 'creation', 'gca', 'stock_program_flag_market',
'collection_stock_program_flag', 'sku_type', 'price_list_eur',
'net_sold_qty_by_coll_rtl_ltd', 'net_sold_qty_by_coll_rtl_ltd_ly',
'net_sold_qty_by_coll_rtl_ltd_lly',
'net_rev_excl_vat_by_coll_rtl_ltd',
'net_rev_excl_vat_by_coll_rtl_ltd__ly',
'net_rev_excl_vat_by_coll_rtl_ltd__lly', 'microcoll', 'received',
'received_ly', 'received_lly', 'key', 'store_code',
'regioncountry', 'sell_tr', 'sell_tr_ly', 'sell_tr_lly',
'cluster', 'price_range', 'colour_group']
```

In [23]:

```
key_col = ['region_desc', 'country', 'store_code', 'sku_code',
'sku_type']

Sell_tr_col = ['sell_tr', 'sell_tr_ly', 'sell_tr_lly']
Received_col = ['received', 'received_ly', 'received_lly']
sold_qty_col = ['net_sold_qty_by_coll_rtl_ltd',
                'net_sold_qty_by_coll_rtl_ltd_ly',
                'net_sold_qty_by_coll_rtl_ltd_lly']

features_col = [
    'price_range', 'colour_group', 'occasion_desc',
    'material_macro_group_desc', 'material_group_desc',
    'dimension_group_desc', 'macro_merc_typology_desc',
    'merchandise_typology_desc', 'macro_line_desc',
    'line_desc', 'model_code', 'model_name'
]
```

qui vengono create delle key_col , per raggruppare tutte quelle variabili utili per il prossimo blocco

In [24]:

```
dupe_cols = df.columns[df.columns.duplicated()]
print("Colonne duplicate:", dupe_cols.tolist())
```

```
Colonne duplicate: ['store_code']
```

In [25]:

```
df = df.loc[:, ~df.columns.duplicated()]
```

ho tolto le colonne duplicate

In [26]:

```
df.shape
```

Out[26]:

```
(68871, 51)
```

In [27]:

```
df.columns
```

Out[27]:

```
Index(['region_desc', 'country', 'channel_desc', 'store_code',
       'store_code_desc', 'entity_typology',
'sku_birth_collection',
       'prod_cat_gender_desc', 'rotb_macro_category_desc',
       'production_category_code', 'prod_category_code_+_desc',
       'macro_merc_typology_desc', 'merchandise_typology_desc',
       'macro_line_desc', 'line_desc', 'model_code', 'model_name',
       'color_desc', 'material_macro_group_desc',
'material_group_desc',
       'dimension_group_desc', 'construction_group',
'occasion_desc',
       'sku_code', 'sku_code_+_desc', 'special_production_desc',
'brand_desc',
       'creation', 'gca', 'stock_program_flag_market',
       'collection_stock_program_flag', 'sku_type',
'price_list_eur',
       'net_sold_qty_by_coll_rtl_ltd',
'net_sold_qty_by_coll_rtl_ltd_ly',
       'net_sold_qty_by_coll_rtl_ltd_lly',
'net_rev_excl_vat_by_coll_rtl_ltd',
       'net_rev_excl_vat_by_coll_rtl_ltd__ly',
       'net_rev_excl_vat_by_coll_rtl_ltd__lly', 'microcoll',
'received',
       'received_ly', 'received_lly', 'key', 'regioncountry',
'sell_tr',
       'sell_tr_ly', 'sell_tr_lly', 'cluster', 'price_range',
'colour_group'],
      dtype='object')
```

In [28]:

```
df.drop_duplicates().shape
```

Out[28]:

```
(68871, 51)
```

In [29]:

```
colour_mapping = {
    'BLACK': 'NEUTRAL', 'WHITE': 'NEUTRAL', 'BEIGE': 'NEUTRAL',
'TAUPE': 'NEUTRAL',
    'GREY': 'NEUTRAL', 'LIGHT GREY': 'NEUTRAL', 'DARK GREY':
'NEUTRAL', 'CAMEL': 'NEUTRAL',
```

```
    'BONE': 'NEUTRAL', 'SILVER': 'NEUTRAL', 'TRANSPARENT':
'NEUTRAL', 'NOT DEFINED': 'NEUTRAL',

    'DARK BROWN': 'BROWN', 'MEDIUM BROWN': 'BROWN', 'TAN':
'BROWN', 'RUST': 'BROWN', 'ORANGE': 'BROWN',

    'RED': 'RED_PINK', 'BORDEAUX': 'RED_PINK', 'LIGHT PINK':
'RED_PINK', 'DARK PINK': 'RED_PINK', 'PURPLE': 'RED_PINK',

    'BLUE': 'BLUE', 'BLUETTE': 'BLUE', 'LIGHT BLUE': 'BLUE',

    'DK OLIVE GREEN': 'GREEN', 'LT OLIVE GREEN': 'GREEN', 'DK
BRILL.GREEN': 'GREEN',
    'LT BRILL.GREEN': 'GREEN', 'DARK AQUA GREEN': 'GREEN', 'LIGHT
AQUA GREEN': 'GREEN',

    'LIGHT YELLOW': 'YELLOW_GOLD', 'DARK YELLOW': 'YELLOW_GOLD',
'GOLD': 'YELLOW_GOLD',

    'LILAC': 'LILAC', 'TURQUOISE': 'TURQUOISE'
}

df['macro_colour_group'] = df['colour_group'].map(colour_mapping)

print(df[['colour_group', 'macro_colour_group']].head(10))
     colour_group macro_colour_group
0      LIGHT PINK           RED_PINK
1           BLACK            NEUTRAL
2           WHITE            NEUTRAL
3  LT OLIVE GREEN              GREEN
4      LIGHT BLUE               BLUE
5           BLACK            NEUTRAL
6             RED           RED_PINK
7  DK OLIVE GREEN              GREEN
8  DK OLIVE GREEN              GREEN
9    MEDIUM BROWN              BROWN
```

per ridurre le variabili di colore e non creare dei valori univoci ho creato dei gruppi su base scale di colori

In [30]:

```
threshold = 500

line_counts = df['macro_line_desc'].value_counts()

rare_lines = line_counts[line_counts < threshold].index

# nuova colonna con categoria 'OTHER'
df['macro_line_grouped'] =
df['macro_line_desc'].replace(rare_lines, 'OTHER')
```

```
print(df['macro_line_grouped'].value_counts())
macro_line_grouped
FERRAGAMO HUG          9248
THE STUDIO             8618
WANDA                  6250
FERRAGAMO CREATION     4332
TRIFOLIO               4033
FERRAGAMO ARCHIVE      4017
*NA                    2885
FIAMMA                 2670
PRISMA                 2378
CUT OUT                2298
MINIBAG                2072
OTHER                  2041
VIVA BOW BAG           1977
RAINBOW MATELASSE'     1753
VARA GROS GRAIN        1538
FERRAGAMO SIGNATURE    1442
GANCINO VELA           1429
GRAPHIC EMBOSSED       1418
FERRAGAMO BEACHWEAR    1394
F.GLAM                 1319
FLORENCE               1275
NEW LINE AI24          1184
TRAVEL                 1139
GANCINI QUILTING        880
STAR                    705
FLAP TRAPEZIO           576
Name: count, dtype: int64
```

anche qui ho deciso di eliminare un po' di ridondanza sulla linea del prodotto

In [31]:

```
(df['received'] == 0).sum()
```

Out[31]:

```
np.int64(43753)
```

in questo notebook proverò ad eliminare le righe che hanno 0 received value, poiché nel 2022 e 2023, ancora segnati dalla lenta ripresa dopo il covid, hanno avuto molto ricevuto ma 0 vendite in quasi l'89% degli store. Ci sono molte righe con 0 ricevuto e quindi direttamente an che con 0 vendite (se non lo hai in negozio non puoi vendere), questo potrebbe fuorviare il modello che potrebbe penalizzare delle caratteristiche di prodotto che nel 2024 invece hanno funzionato molto bene

In [32]:

```
df = df[df['received'] != 0].copy()
```

In [33]:

```
df.shape
```

Out[33]:

```
(25118, 53)
```

In [34]:

```
df = df[df['net_sold_qty_by_coll_rtl_ltd'] <=
df['received']].copy()
```

In [35]:

```
df.columns.tolist()
```

Out[35]:

```
['region_desc',
 'country',
 'channel_desc',
 'store_code',
 'store_code_desc',
 'entity_typology',
 'sku_birth_collection',
 'prod_cat_gender_desc',
 'rotb_macro_category_desc',
 'production_category_code',
 'prod_category_code_+_desc',
 'macro_merc_typology_desc',
 'merchandise_typology_desc',
 'macro_line_desc',
 'line_desc',
 'model_code',
 'model_name',
 'color_desc',
 'material_macro_group_desc',
 'material_group_desc',
 'dimension_group_desc',
 'construction_group',
 'occasion_desc',
 'sku_code',
 'sku_code_+_desc',
 'special_production_desc',
 'brand_desc',
 'creation',
 'gca',
 'stock_program_flag_market',
 'collection_stock_program_flag',
 'sku_type',
 'price_list_eur',
 'net_sold_qty_by_coll_rtl_ltd',
 'net_sold_qty_by_coll_rtl_ltd_ly',
 'net_sold_qty_by_coll_rtl_ltd_lly',
 'net_rev_excl_vat_by_coll_rtl_ltd',
 'net_rev_excl_vat_by_coll_rtl_ltd__ly',
 'net_rev_excl_vat_by_coll_rtl_ltd__lly',
 'microcoll',
 'received',
 'received_ly',
 'received_lly',
 'key',
 'regioncountry',
```

```
 'sell_tr',
 'sell_tr_ly',
 'sell_tr_lly',
 'cluster',
 'price_range',
 'colour_group',
 'macro_colour_group',
 'macro_line_grouped']
```

In [36]:

```
df.rename(columns={
    'net_sold_qty_by_coll_rtl_ltd': 'sold_qty',
    'net_sold_qty_by_coll_rtl_ltd_ly': 'sold_qty_ly',
    'net_sold_qty_by_coll_rtl_ltd_lly': 'sold_qty_lly'
}, inplace=True)
```

In [37]:

```
df.columns.tolist()
```

Out[37]:

```
['region_desc',
 'country',
 'channel_desc',
 'store_code',
 'store_code_desc',
 'entity_typology',
 'sku_birth_collection',
 'prod_cat_gender_desc',
 'rotb_macro_category_desc',
 'production_category_code',
 'prod_category_code_+_desc',
 'macro_merc_typology_desc',
 'merchandise_typology_desc',
 'macro_line_desc',
 'line_desc',
 'model_code',
 'model_name',
 'color_desc',
 'material_macro_group_desc',
 'material_group_desc',
 'dimension_group_desc',
 'construction_group',
 'occasion_desc',
 'sku_code',
 'sku_code_+_desc',
 'special_production_desc',
 'brand_desc',
 'creation',
 'gca',
 'stock_program_flag_market',
 'collection_stock_program_flag',
 'sku_type',
 'price_list_eur',
 'sold_qty',
```

```
 'sold_qty_ly',
 'sold_qty_lly',
 'net_rev_excl_vat_by_coll_rtl_ltd',
 'net_rev_excl_vat_by_coll_rtl_ltd__ly',
 'net_rev_excl_vat_by_coll_rtl_ltd__lly',
 'microcoll',
 'received',
 'received_ly',
 'received_lly',
 'key',
 'regioncountry',
 'sell_tr',
 'sell_tr_ly',
 'sell_tr_lly',
 'cluster',
 'price_range',
 'colour_group',
 'macro_colour_group',
 'macro_line_grouped']
```

i dati ora sono stati ripuliti e pronti per essere utilizzati per lo studio sul modello

In [38]:

```
df.to_csv("df_clean_3.csv", index=False)
```

PROVA MODELLI

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
pd.set_option('display.max_columns', 500)
```

In [2]:

```
df =
pd.read_csv("/Users/stefanolandolfi/Desktop/FERRAGAMO_PYTHON/01_py
thon/df_clean_3.csv")
df.head()
```

Out[2]:

esplorazione dataset per vedere risposta di sold_qty sulle variabili

In [3]:

```
df.shape
```

Out[3]:

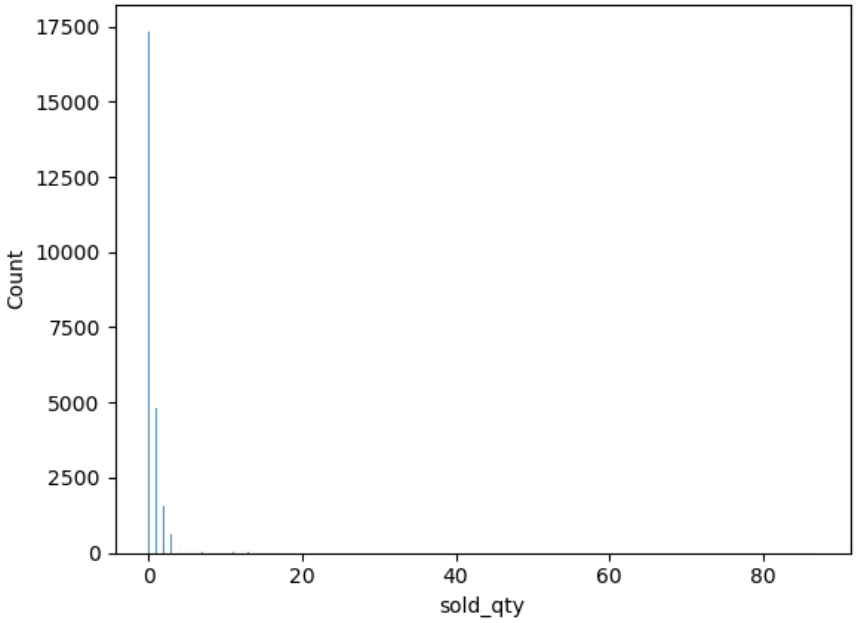
```
(25117, 53)
```

In [4]:

```
sns.histplot(data=df, x='sold_qty')
```

Out[4]:

```
<Axes: xlabel='sold_qty', ylabel='Count'>
```

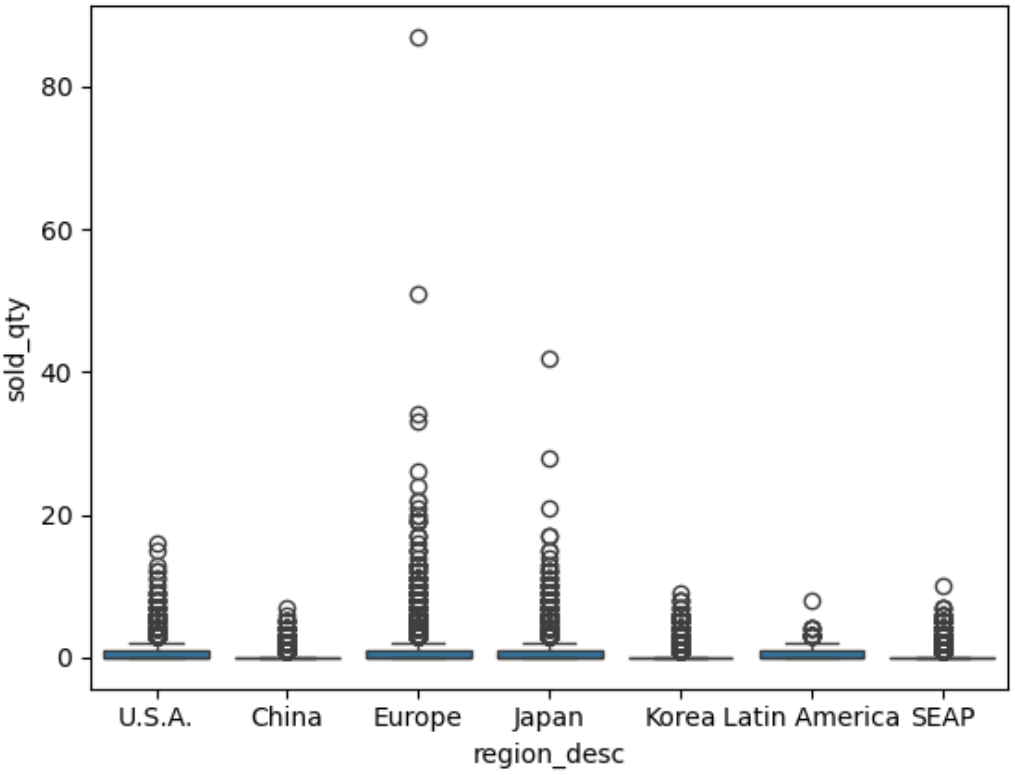


In [5]:

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.boxplot(data=df, x='region_desc', y='sold_qty')
```

Out[5]:

```
<Axes: xlabel='region_desc', ylabel='sold_qty'>
```



nel prossimo codice ho esplorato come si comporta sold_qty nelle varie region. vedendo media , mediana e std. nella parte finale c'è la percentuale degli sku che hanno veduto

In [6]:

```
region_summary = df.groupby("region_desc").agg(
    n_obs=("sold_qty", "count"),
    sold_qty_mean=("sold_qty", "mean"),
    sold_qty_median=("sold_qty", "median"),
    sold_qty_std=("sold_qty", "std"),
    sold_qty_positive_pct=("sold_qty", lambda x: (x > 0).mean())
).sort_values(by="sold_qty_mean", ascending=False)

region_summary.reset_index(inplace=True)
region_summary
```

Out[6]:

	region_desc	n_obs	sold_qty_mean	sold_qty_median	sold_qty_std	sold_qty_positive_pct
0	Europe	4196	1.174690	0.0	2.721869	0.486892
1	Japan	3017	0.951608	0.0	1.954636	0.432217
2	U.S.A.	4724	0.671677	0.0	1.208321	0.385478
3	Latin America	1551	0.386202	0.0	0.685172	0.303030
4	SEAP	3517	0.352005	0.0	0.805664	0.231447
5	Korea	3530	0.324646	0.0	0.849042	0.196601
6	China	4582	0.178306	0.0	0.532572	0.134439

In [7]:

```
df.groupby('region_desc')['sold_qty'].describe()
```

Out[7]:

	count	mean	std	min	25%	50%	75%	max
region_desc								
China	4582.0	0.178306	0.532572	0.0	0.0	0.0	0.0	7.0
Europe	4196.0	1.174690	2.721869	0.0	0.0	0.0	1.0	87.0
Japan	3017.0	0.951608	1.954636	0.0	0.0	0.0	1.0	42.0
Korea	3530.0	0.324646	0.849042	0.0	0.0	0.0	0.0	9.0
Latin America	1551.0	0.386202	0.685172	0.0	0.0	0.0	1.0	8.0
SEAP	3517.0	0.352005	0.805664	0.0	0.0	0.0	0.0	10.0
U.S.A.	4724.0	0.671677	1.208321	0.0	0.0	0.0	1.0	16.0

si può notare come il 34% dei prodotti ha venduto in america e le sold_qty variano molto al variare di ogni region

In [8]:

```
country_summary = df.groupby("country").agg(
    n_obs=("sold_qty", "count"),
    sold_qty_mean=("sold_qty", "mean"),
    sold_qty_median=("sold_qty", "median"),
    sold_qty_std=("sold_qty", "std"),
    sold_qty_positive_pct=("sold_qty", lambda x: (x > 0).mean())
).sort_values(by="sold_qty_mean", ascending=False)
```

```
country_summary.reset_index(inplace=True)
country_summary
```

Out[8]:

	country	n_obs	sold_qty_mean	sold_qty_median	sold_qty_std	sold_qty_positive_pct
0	Italy	1681	1.713861	1.0	3.781012	0.556811
1	Argentina	46	1.347826	1.0	1.369791	0.782609
2	Austria	187	1.187166	1.0	1.870037	0.540107
3	Belgium	96	1.020833	1.0	1.673189	0.552083
4	France	550	0.980000	0.0	1.986649	0.480000
5	Japan	3017	0.951608	0.0	1.954636	0.432217
6	United Kingdom	385	0.844156	0.0	1.615879	0.438961
7	Spain	468	0.829060	0.0	1.465028	0.470085
8	Principality of Monaco	100	0.700000	0.0	1.487320	0.350000
9	United States of America	4402	0.691958	0.0	1.234059	0.392322
10	Switzerland	237	0.679325	0.0	1.210177	0.413502
11	Chile	96	0.593750	0.0	0.828164	0.406250
12	Germany	492	0.497967	0.0	1.002033	0.339431
13	Taiwan	1042	0.427063	0.0	0.942617	0.265835
14	Singapore	395	0.410127	0.0	0.827140	0.265823
15	Canada	322	0.394410	0.0	0.721089	0.291925
16	Australia	645	0.378295	0.0	0.779887	0.260465
17	Hong Kong	669	0.355755	0.0	0.784706	0.231689
18	Mexico	1409	0.340667	0.0	0.612651	0.280341
19	South Korea	3530	0.324646	0.0	0.849042	0.196601
20	Malaysia	236	0.254237	0.0	0.661223	0.182203
21	Thailand	232	0.206897	0.0	0.658179	0.142241

	country	n_obs	sold_qty_mean	sold_qty_median	sold_qty_std	sold_qty_positive_pct
22	China	4582	0.178306	0.0	0.532572	0.134439
23	Macau	298	0.137584	0.0	0.423863	0.110738

In [9]:

```
sku_positive_count = df[df['sold_qty'] >
0].groupby('country')['sku_code'].nunique().sort_values(ascending=
False)
sku_positive_count.reset_index(name='n_sku_sold_positive')
```

Out[9]:

	country	n_sku_sold_positive
0	United States of America	252
1	Italy	232
2	Japan	166
3	France	148
4	Spain	133
5	China	131
6	United Kingdom	131
7	Mexico	125
8	South Korea	121
9	Taiwan	115
10	Austria	101
11	Germany	98
12	Hong Kong	90
13	Australia	81
14	Switzerland	75
15	Canada	63
16	Singapore	61
17	Belgium	53
18	Chile	39
19	Argentina	36
20	Principality of Monaco	35
21	Malaysia	30
22	Macau	28
23	Thailand	24

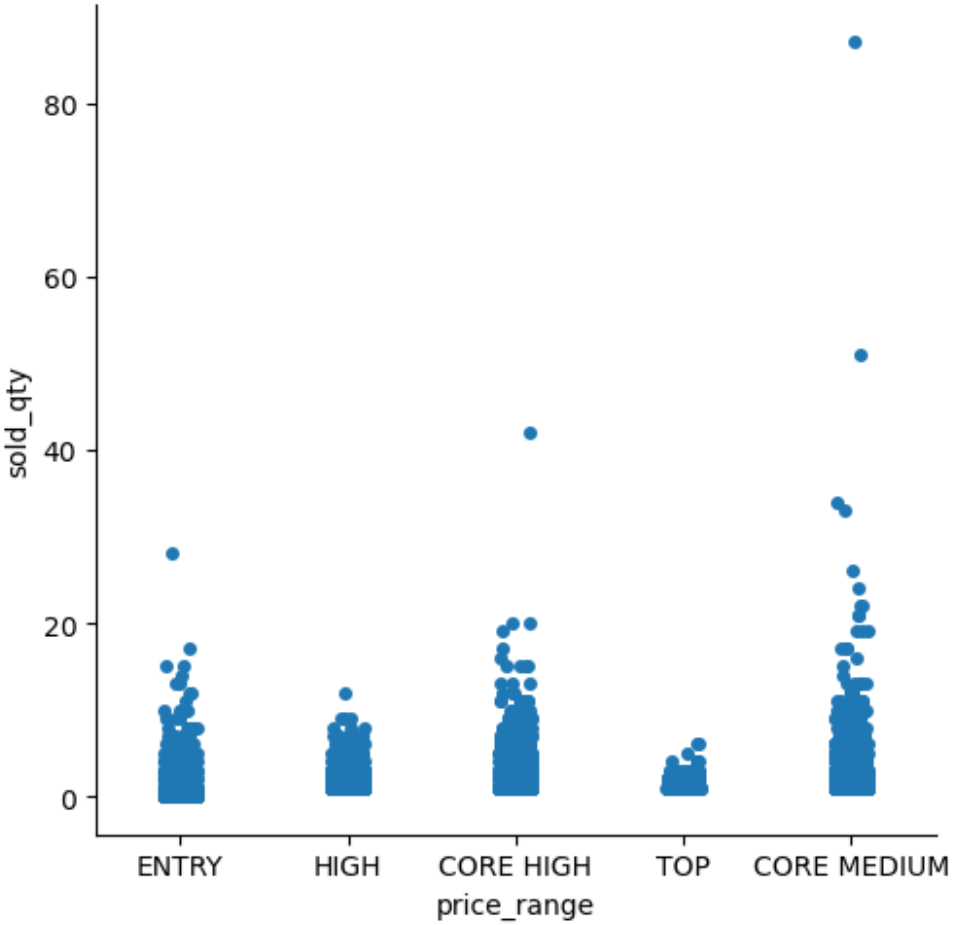
anche le vendite nelle country sembrano cambiare molto, risultato inatteso, evitiamo lo store_code poichè è gia parte della country

In [10]:

```
sns.catplot(data=df, x='price_range', y='sold_qty')
```

Out[10]:

```
<seaborn.axisgrid.FacetGrid at 0x15a92deb0>
```



In [11]:

```
sku_positive = df[df['sold_qty'] >
0].groupby('price_range')['sku_code'].nunique().sort_values(ascend
ing=False)
```

In [12]:

```
sku_positive.reset_index(name='n_sku_sold_positive_price')
```

Out[12]:

	price_range	n_sku_sold_positive_price
0	HIGH	267
1	CORE HIGH	234
2	CORE MEDIUM	140
3	TOP	111
4	ENTRY	66

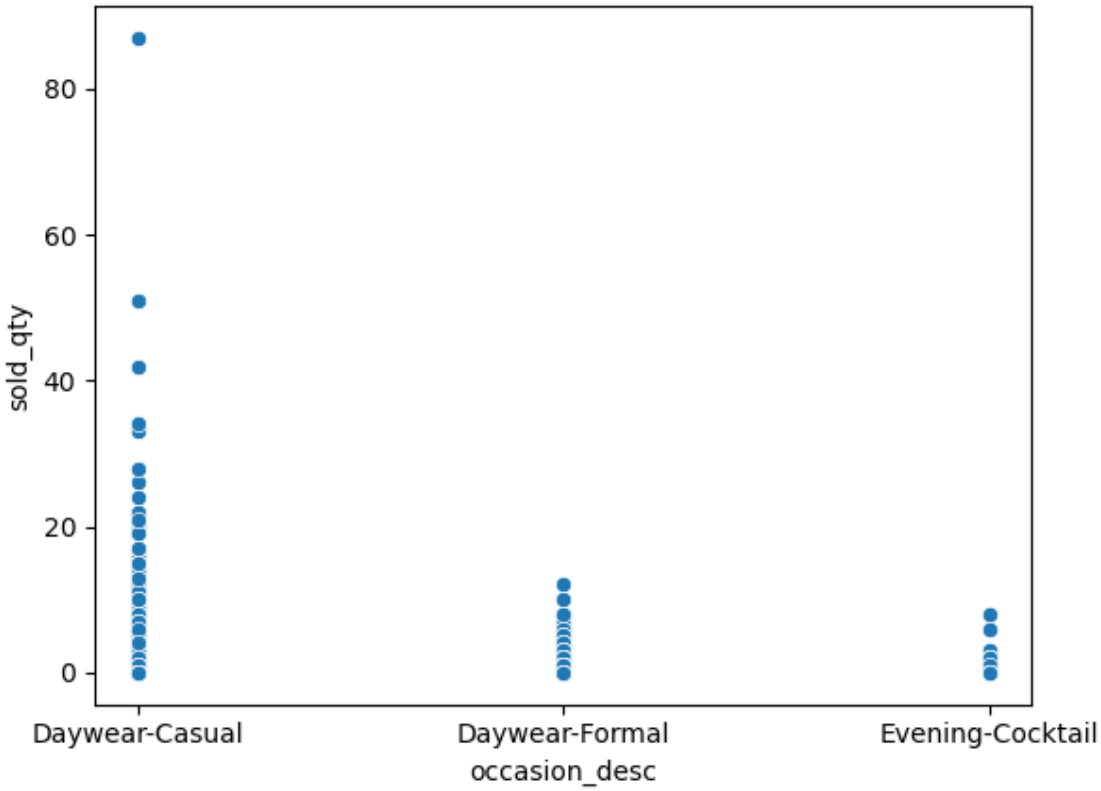
colour group non verrà analizzato, poichè nella parte di preparazione dei dati è stato creato un macro colore per ogni gruppo in base a scalanature di colore , poichè c'è il rischio che vengano inseriti sempre nuovi tipi di colore e soprattutto presentava troppe variabili univoche

In [13]:

```
sns.scatterplot(data=df, x='occasion_desc', y='sold_qty')
```

Out[13]:

```
<Axes: xlabel='occasion_desc', ylabel='sold_qty'>
```



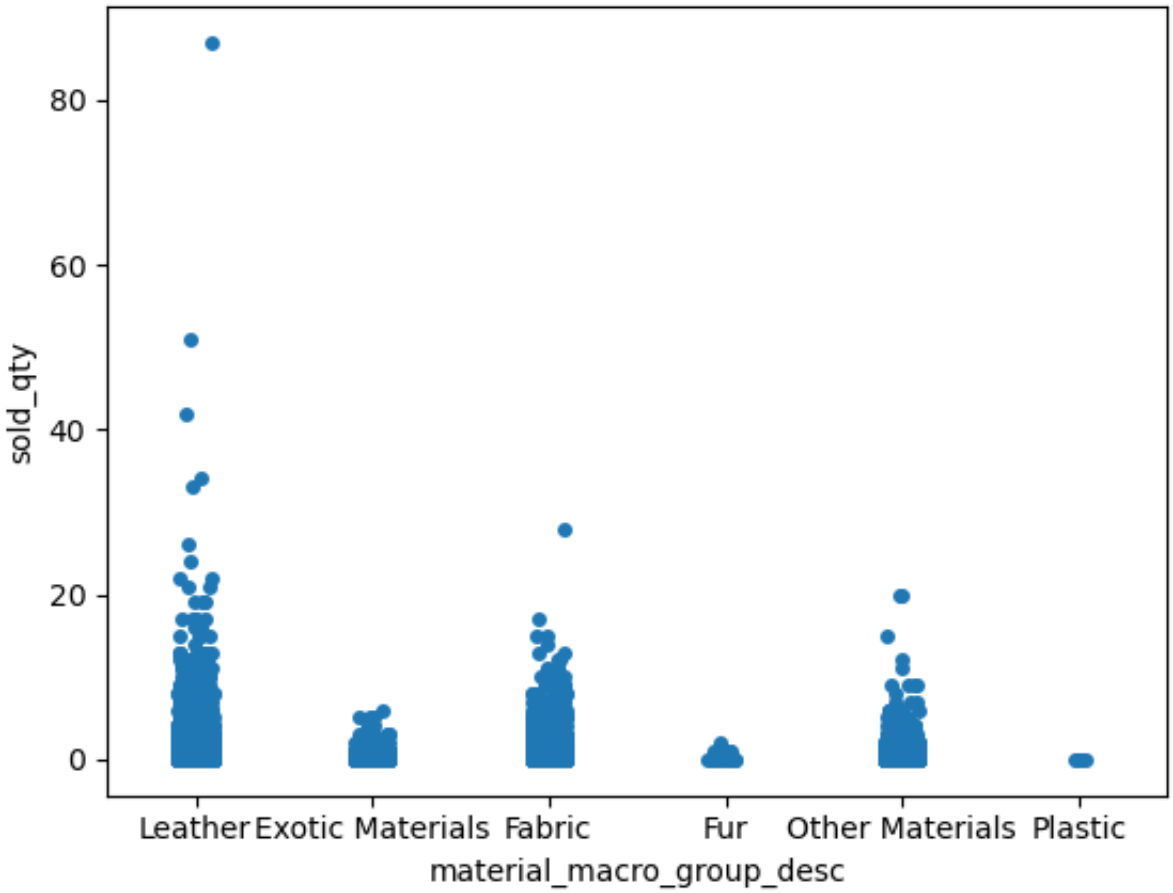
solo tre variabili , gli sku sono troppo raccolti, non influiscono sulle vendite

In [14]:

```
sns.stripplot(data=df, x='material_macro_group_desc',
y='sold_qty')
```

Out[14]:

```
<Axes: xlabel='material_macro_group_desc', ylabel='sold_qty'>
```



In [15]:

```
sku_positive_mat = df[df['sold_qty'] >
0].groupby('material_macro_group_desc')['sku_code'].nunique().sort
_values(ascending=False)
sku_positive_mat.reset_index(name='n_sku_sold_positive_mat')
```

Out[15]:

	material_macro_group_desc	n_sku_sold_positive_mat
0	Leather	312
1	Exotic Materials	56
2	Fabric	33
3	Other Materials	19
4	Fur	2

material_macro_group come si nota presenta troppe variabili con pochissime righe, se escludiamo queste variabili ne rimangono 3 di cui 1 prevale su tutte che la pelle , meglio non includere potrebbe deviare il modello dalla considerazione

In [16]:

```
sku_positive_mate= df[df['sold_qty'] >
0].groupby('material_group_desc')['sku_code'].nunique().sort_value
s(ascending=False)
sku_positive_mate.reset_index(name='n_sku_sold_positive_mate')
```

Out[16]:

	material_group_desc	n_sku_sold_positive_mate
0	CALF	288
1	FABRIC	31
2	OSTRICH	26
3	LS-INTRECCIO	12
4	KID	9
5	ALLIGATOR	7
6	LS - Vario a pezzi	7
7	REPTILE	6
8	PYTON	6
9	FISH	5
10	LS-APPL.BORCHIE/s	5
11	CROCODILE	4
12	LIZARD	3
13	BUC SUEDE CALF	3
14	LS-RICAMO	2
15	SKIN LNG	2
16	LS-LAVORAZ.MAGLI	2
17	FEATHER	2
18	LEATHER	1
19	LS - Stampa Generica	1

Calf avrebbe troppo distacco e sono ininfluenti nel modello , anche raggruppando la situazione non cambierebbe

In [17]:

```
sku_positive_dimension_group_desc= df[df['sold_qty'] >
0].groupby('dimension_group_desc')['sku_code'].nunique().sort_valu
es(ascending=False)
sku_positive_dimension_group_desc.reset_index(name='n_sku_sold_pos
itive_dimension_group_desc')
```

Out[17]:

	dimension_group_desc	n_sku_sold_positive_dimension_group_desc
0	Small	207
1	Medium	101
2	Large	59
3	Mini	55

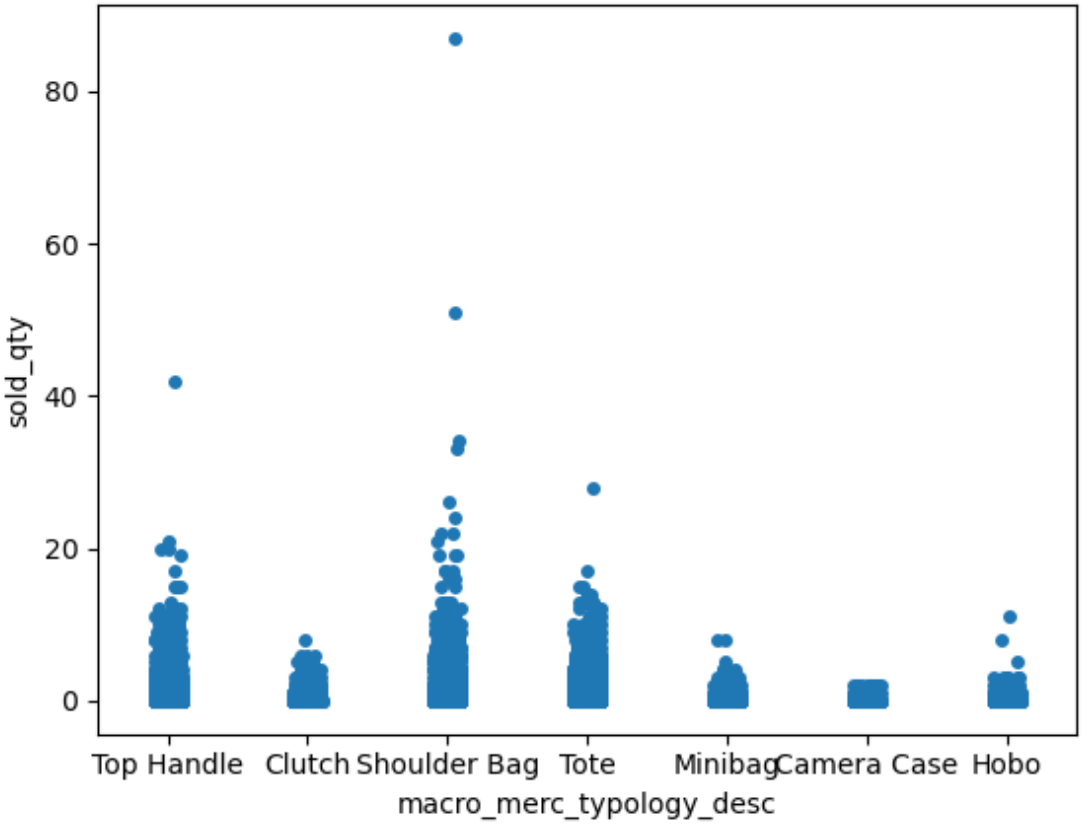
In [18]:

```
# Conteggio delle righe per ogni categoria della variabile
'dimension_group_desc'
dimension_counts =
df['dimension_group_desc'].value_counts().reset_index()
dimension_counts.columns = ['dimension_group_desc', 'n_rows']
```

```
print(dimension_counts)
  dimension_group_desc  n_rows
0                Small   13544
1               Medium    5346
2                 Mini    3861
3                Large    2363
4          Extra Large       3
```

La dimension va inserita, come si può notare le small sono le più prodotte, su un numero elevato di produzione sicuramente avranno molte vendite. Le percentuali di vendita su ogni dimensione vs tot della produzione rimane stabile per ognuna

In [19]:
```
sns.stripplot(data=df, x='macro_merc_typology_desc', y='sold_qty')
```
Out[19]:
```
<Axes: xlabel='macro_merc_typology_desc', ylabel='sold_qty'>
```



In [20]:
```
sku_positive_macro_merc_typology_desc= df[df['sold_qty'] >
0].groupby('macro_merc_typology_desc')['sku_code'].nunique().sort_
values(ascending=False)
sku_positive_macro_merc_typology_desc.reset_index(name='n_sku_sold
_positive_macro_merc_typology_desc')
```
Out[20]:

	macro_merc_typology_desc	n_sku_sold_positive_macro_merc_typology_desc
0	Top Handle	244

	macro_merc_typology_desc	n_sku_sold_positive_macro_merc_typology_desc
1	Tote	61
2	Shoulder Bag	55
3	Clutch	40
4	Minibag	14
5	Hobo	5
6	Camera Case	3

elimino le righe con backpack e luggage

In [21]:
```
df = df[~df['macro_merc_typology_desc'].isin(['Luggage',
'Backpack'])]
```

In [22]:
```
sku_positive_macro_merc_typology_desc= df[df['sold_qty'] >
0].groupby('macro_merc_typology_desc')['sku_code'].nunique().sort_
values(ascending=False)
sku_positive_macro_merc_typology_desc.reset_index(name='n_sku_sold
_positive_macro_merc_typology_desc')
```

Out[22]:

	macro_merc_typology_desc	n_sku_sold_positive_macro_merc_typology_desc
0	Top Handle	244
1	Tote	61
2	Shoulder Bag	55
3	Clutch	40
4	Minibag	14
5	Hobo	5
6	Camera Case	3

ora la macromerch puo essere una feature con una distribuzione adeguata

In [23]:
```
sku_positive_merchandise_typology_desc= df[df['sold_qty'] >
0].groupby('merchandise_typology_desc')['sku_code'].nunique().sort
_values(ascending=False)
sku_positive_merchandise_typology_desc.reset_index(name='n_sku_sol
d_positive_merchandise_typology_desc')
```

Out[23]:

	merchandise_typology_desc	n_sku_sold_positive_merchandise_typology_desc
0	Top Handle	241
1	Tote - Handheld	61
2	Clutch	40

	merchandise_typology_desc	n_sku_sold_positive_merchandise_typology_desc
3	Flap	40
4	Minibag	14
5	Shoulder Bag	9
6	Crossbody	6
7	Hobo	5
8	Bowling Bag	3
9	Camera Case	3

La merchandise_typology non va inserita poichè inserisce solo variabili univoche a ciò che già c'è nella macro merch

Per le variabili di prodotto serve un discorso molto ampio per linea,macro linea ecc.... Queste caratteristiche hanno molte variabile, quindi potrebbero essere utili, ma in chiave futura sono variabili che vengono aggiunte nel tempo. ad esempio, ora Ferragamo ha fatto uscire la soft bag con una nuova macro-linea il modello non riuscirebbe ad interpretarla adeguatamente, ma comunque si proverà ad inserire per un discorso di forma della borsa, in quanto Ferragamo non ha nel database una variabile con le forme. Inserirla nel modello potrebbe indirizzare verso un tipo di borsa.

quindi nell'analisi svolta : Region, Country, price_range, macro_colour group, Macro_line_grouped, dimension_group_desc macro_merc_typology_desc

sono le variabili che potrebbero andare nel modello

Tutti i coefficienti sulle regioni e country sono esageratamente alti o vicini a zero, con z-score bassissimi e p-value > 0.05. Probabile problema numerico di collinearità o dati troppo sbilanciati. Molti coefficienti tipo 7.064e+10 indicano che alcune variabili sono quasi collineari → potremmo dover ridurre i dummies.

In [24]:
```
df['vendente'] = (df['sold_qty'] > 0).astype(int)
```

In [25]:
```
feature_cols = [
    'region_desc',
    'country',
    'price_range',
    'macro_colour_group',
    'macro_merc_typology_desc',
    'dimension_group_desc',
    'macro_line_grouped',
    'store_code'
]


df_model = df[feature_cols + ['vendente']].dropna()
```

In [26]:

```
X = pd.get_dummies(df_model.drop(columns='vendente'),
drop_first=True)
y = df_model['vendente']
```

In [27]:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report,
confusion_matrix, roc_auc_score

# Split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Modello logistico
clf = LogisticRegression(max_iter=1000)
clf.fit(X_train, y_train)

# Predizioni
y_pred = clf.predict(X_test)
y_proba = clf.predict_proba(X_test)[:, 1]
```

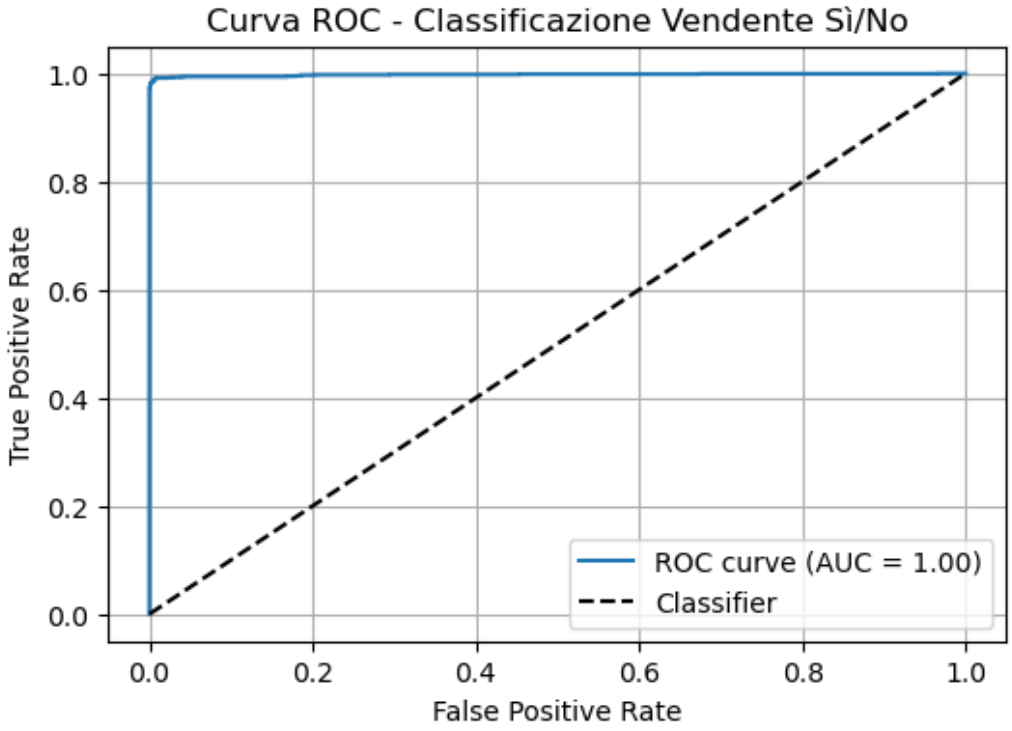
In [28]:

```
print("Accuracy:", clf.score(X_test, y_test))
print("ROC AUC:", roc_auc_score(y_test, y_proba))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("Classification Report:\n", classification_report(y_test,
y_pred))
Accuracy: 0.9930334394904459
ROC AUC: 0.9977262191915045
Confusion Matrix:
 [[3465    8]
 [  27 1524]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      3473
           1       0.99      0.98      0.99      1551

    accuracy                           0.99      5024
   macro avg       0.99      0.99      0.99      5024
weighted avg       0.99      0.99      0.99      5024
```

In [29]:

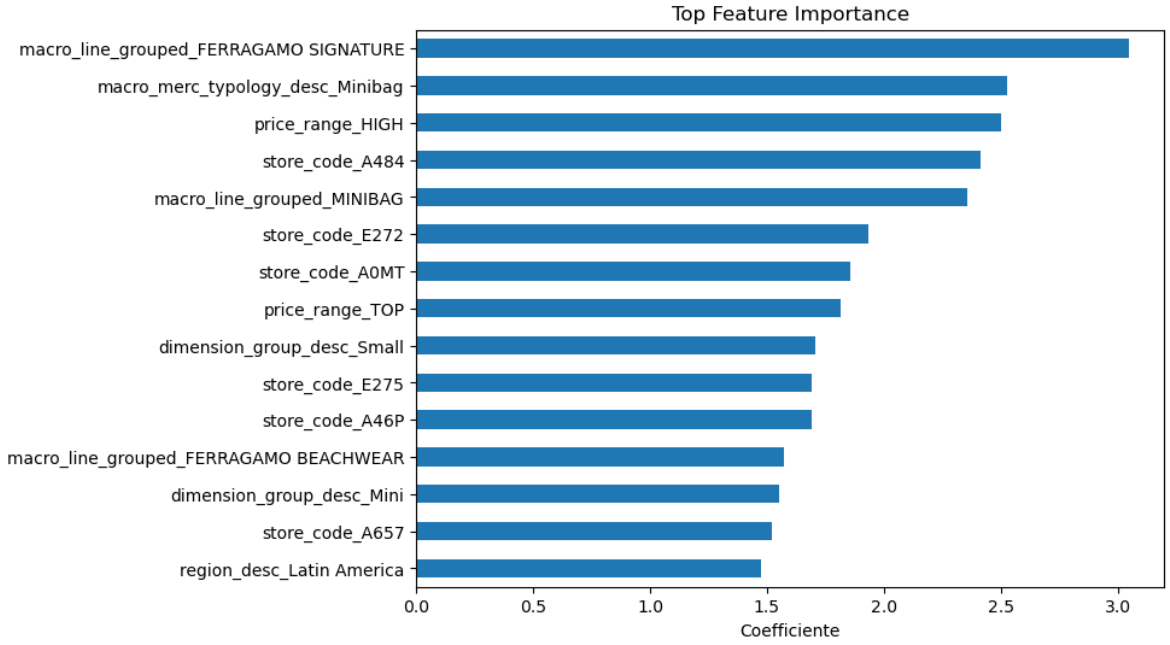
```
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve

fpr, tpr, thresholds = roc_curve(y_test, y_proba)
plt.figure(figsize=(6,4))
plt.plot(fpr, tpr, label=f'ROC curve (AUC = {roc_auc_score(y_test,
y_proba):.2f})')
plt.plot([0, 1], [0, 1], 'k--', label='Classifier')
```

```
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Curva ROC - Classificazione Vendente Sì/No')
plt.legend()
plt.grid()
plt.show()
```



In [30]:

```
feature_importance = pd.Series(clf.coef_[0], index=X.columns)
feature_importance.sort_values(ascending=False).head(15).plot(kind
='barh', figsize=(8,6), title="Top Feature Importance")
plt.xlabel("Coefficiente")
plt.gca().invert_yaxis()
plt.show()
```



ora poisson

In [31]:

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
```

In [32]:

```
df_poisson = df[(df["sold_qty"] > 0) & (df["sold_qty"] <=
20)].copy()
```

In [33]:

```
formula = """sold_qty ~ C(region_desc) + C(country) +
             C(price_range) + C(macro_colour_group) +
             C(macro_merc_typology_desc) + C(dimension_group_desc)
+ C(macro_line_grouped) + C(store_code)"""
```

In [34]:

```
poisson_model = smf.glm(formula=formula,
                        data=df_poisson,
                        family=sm.families.Poisson()).fit()
```

In [35]:

```
print(poisson_model.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
============
Dep. Variable:                sold_qty   No. Observations:
7750
Model:                            GLM   Df Residuals:
7432
Model Family:                 Poisson   Df Model:
317
Link Function:                    Log   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-11679.
Date:                Fri, 19 Sep 2025   Deviance:
5077.9
Time:                        19:48:20   Pearson chi2:
6.04e+03
No. Iterations:                   100   Pseudo R-squ. (CS):
0.3053
Covariance Type:            nonrobust
==============================================================================
================================================
                                                    coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------
Intercept                                  -1.059e+12   1.74e+12
-0.607      0.544   -4.48e+12    2.36e+12
```

```
C(region_desc)[T.Europe]                      1.091e+12    1.8e+12
0.607      0.544   -2.43e+12    4.61e+12
C(region_desc)[T.Japan]                       7.287e+11    1.2e+12
0.607      0.544   -1.63e+12    3.08e+12
C(region_desc)[T.Korea]                       5.979e+11   9.85e+11
0.607      0.544   -1.33e+12    2.53e+12
C(region_desc)[T.Latin America]               8.248e+11   1.36e+12
0.607      0.544   -1.84e+12    3.49e+12
C(region_desc)[T.SEAP]                        8.914e+11   1.47e+12
0.607      0.544   -1.99e+12    3.77e+12
C(region_desc)[T.U.S.A.]                      4.836e+11   7.97e+11
0.607      0.544   -1.08e+12    2.05e+12
C(country)[T.Australia]                       1.671e+11   2.75e+11
0.607      0.544   -3.73e+11    7.07e+11
C(country)[T.Austria]                        -6.768e+10   1.12e+11
-0.607      0.544   -2.86e+11    1.51e+11
C(country)[T.Belgium]                         -8.75e+10   1.44e+11
-0.607      0.544    -3.7e+11    1.95e+11
C(country)[T.Canada]                          5.352e+11   8.82e+11
0.607      0.544   -1.19e+12    2.26e+12
C(country)[T.Chile]                           5.646e+10   9.31e+10
0.607      0.544   -1.26e+11    2.39e+11
C(country)[T.China]                           1.042e+12   1.72e+12
0.607      0.544   -2.32e+12    4.41e+12
C(country)[T.France]                         -7.228e+10   1.19e+11
-0.607      0.544   -3.06e+11    1.61e+11
C(country)[T.Germany]                           3.1e+10   5.11e+10
0.607      0.544   -6.91e+10    1.31e+11
C(country)[T.Hong Kong]                       1.256e+11   2.07e+11
0.607      0.544    -2.8e+11    5.31e+11
C(country)[T.Italy]                           2.547e+10    4.2e+10
0.607      0.544   -5.68e+10    1.08e+11
C(country)[T.Japan]                           3.209e+11   5.29e+11
0.607      0.544   -7.16e+11    1.36e+12
C(country)[T.Macau]                           9.439e+10   1.56e+11
0.607      0.544   -2.11e+11    3.99e+11
C(country)[T.Malaysia]                         1.91e+11   3.15e+11
0.607      0.544   -4.26e+11    8.08e+11
C(country)[T.Mexico]                           2.08e+11   3.43e+11
0.607      0.544   -4.64e+11     8.8e+11
C(country)[T.Principality of Monaco]          3.427e+10   5.65e+10
0.607      0.544   -7.64e+10    1.45e+11
C(country)[T.Singapore]                       1.728e+11   2.85e+11
0.607      0.544   -3.85e+11    7.31e+11
C(country)[T.South Korea]                     4.525e+11   7.46e+11
0.607      0.544   -1.01e+12    1.91e+12
C(country)[T.Spain]                          -3.043e+10   5.02e+10
-0.607      0.544   -1.29e+11    6.79e+10
C(country)[T.Switzerland]                    -5.291e+10   8.72e+10
-0.607      0.544   -2.24e+11    1.18e+11
```

```
C(country)[T.Taiwan]                              1.594e+11   2.63e+11
0.607      0.544   -3.55e+11    6.74e+11
C(country)[T.Thailand]                             1.07e+11   1.76e+11
0.607      0.544   -2.39e+11    4.52e+11
C(country)[T.United Kingdom]                  -2.758e+10   4.55e+10
-0.607      0.544   -1.17e+11    6.15e+10
C(country)[T.United States of America]        5.625e+11   9.27e+11
0.607      0.544   -1.25e+12    2.38e+12
C(price_range)[T.CORE MEDIUM]                    -0.0132      0.030
-0.446      0.655      -0.071       0.045
C(price_range)[T.ENTRY]                           0.0846      0.053
1.611      0.107      -0.018       0.188
C(price_range)[T.HIGH]                           -0.1256      0.028
-4.486      0.000      -0.181      -0.071
C(price_range)[T.TOP]                            -0.2372      0.054
-4.430      0.000      -0.342      -0.132
C(macro_colour_group)[T.BROWN]                    0.2792      0.035
8.014      0.000       0.211       0.348
C(macro_colour_group)[T.GREEN]                    0.0115      0.039
0.297      0.767      -0.065       0.088
C(macro_colour_group)[T.LILAC]                   -0.1174      0.052
-2.243      0.025      -0.220      -0.015
C(macro_colour_group)[T.NEUTRAL]                  0.0513      0.026
1.977      0.048       0.000       0.102
C(macro_colour_group)[T.RED_PINK]                 0.0654      0.041
1.576      0.115      -0.016       0.147
C(macro_colour_group)[T.TURQUOISE]               -0.3031      0.416
-0.729      0.466      -1.119       0.512
C(macro_colour_group)[T.YELLOW_GOLD]             -0.0341      0.113
-0.301      0.763      -0.256       0.188
C(macro_merc_typology_desc)[T.Clutch]            -0.2383      0.190
-1.257      0.209      -0.610       0.133
C(macro_merc_typology_desc)[T.Hobo]               0.1850      0.204
0.906      0.365      -0.215       0.585
C(macro_merc_typology_desc)[T.Minibag]            0.1410      0.173
0.816      0.414      -0.198       0.480
C(macro_merc_typology_desc)[T.Shoulder Bag]       0.3355      0.182
1.848      0.065      -0.020       0.691
C(macro_merc_typology_desc)[T.Top Handle]        -0.0282      0.185
-0.152      0.879      -0.391       0.335
C(macro_merc_typology_desc)[T.Tote]               0.3634      0.184
1.980      0.048       0.004       0.723
C(dimension_group_desc)[T.Medium]                 0.2881      0.048
6.004      0.000       0.194       0.382
C(dimension_group_desc)[T.Mini]                   0.6491      0.059
11.081      0.000       0.534       0.764
C(dimension_group_desc)[T.Small]                  0.4771      0.048
9.959      0.000       0.383       0.571
C(macro_line_grouped)[T.CUT OUT]                  0.0979      0.190
0.515      0.606      -0.275       0.470
```

```
C(macro_line_grouped)[T.FERRAGAMO ARCHIVE]      -0.1353      0.164
-0.823      0.411      -0.458       0.187
C(macro_line_grouped)[T.FERRAGAMO BEACHWEAR]    -0.1513      0.188
-0.804      0.421      -0.520       0.217
C(macro_line_grouped)[T.FERRAGAMO CREATION]      0.0058      0.122
0.048      0.962      -0.234       0.245
C(macro_line_grouped)[T.FERRAGAMO HUG]           0.4979      0.111
4.492      0.000       0.281       0.715
C(macro_line_grouped)[T.FERRAGAMO SIGNATURE]     0.4450      0.126
3.544      0.000       0.199       0.691
C(macro_line_grouped)[T.FIAMMA]                 -0.2435      0.122
-1.999      0.046      -0.482      -0.005
C(macro_line_grouped)[T.FLORENCE]                0.0838      0.129
0.650      0.516      -0.169       0.337
C(macro_line_grouped)[T.GANCINO VELA]            0.9151      0.138
6.636      0.000       0.645       1.185
C(macro_line_grouped)[T.MINIBAG]                 0.2699      0.275
0.980      0.327      -0.270       0.809
C(macro_line_grouped)[T.NEW LINE AI24]           0.3881      0.119
3.271      0.001       0.156       0.621
C(macro_line_grouped)[T.OTHER]                  -0.0666      0.137
-0.487      0.626      -0.334       0.201
C(macro_line_grouped)[T.PRISMA]                  0.1365      0.136
1.007      0.314      -0.129       0.402
C(macro_line_grouped)[T.RAINBOW MATELASSE']      0.0712      0.120
0.594      0.553      -0.164       0.306
C(macro_line_grouped)[T.STAR]                    0.2047      0.131
1.568      0.117      -0.051       0.461
C(macro_line_grouped)[T.THE STUDIO]              0.3065      0.113
2.714      0.007       0.085       0.528
C(macro_line_grouped)[T.VARA GROS GRAIN]         0.0565      0.295
0.191      0.848      -0.523       0.636
C(macro_line_grouped)[T.WANDA]                  -0.2299      0.172
-1.338      0.181      -0.567       0.107
C(store_code)[T.A0AU]                         8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A0AW]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0AX]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0AY]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0AZ]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0BG]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0BO]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0BP]                         1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
```

```
C(store_code)[T.A0ED]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A0EQ]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0ER]                          7.274e+10    1.2e+11
0.607      0.544   -1.62e+11    3.08e+11
C(store_code)[T.A0FZ]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A0GE]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A0GI]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0IM]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0IO]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0KU]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0KV]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A0MP]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A0MT]                            -0.1360      0.250
-0.544      0.587      -0.626       0.354
C(store_code)[T.A0NC]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0OQ]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A0PN]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0PO]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0PP]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0QE]                          7.274e+10    1.2e+11
0.607      0.544   -1.62e+11    3.08e+11
C(store_code)[T.A0TA]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0TU]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A0WC]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A0XC]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0XQ]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0YQ]                           4.15e+10   6.84e+10
0.607      0.544   -9.26e+10    1.76e+11
C(store_code)[T.A0ZL]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
```

```
C(store_code)[T.A0ZR]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0ZS]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A0ZZ]                          6.018e+10   9.92e+10
0.607      0.544   -1.34e+11    2.55e+11
C(store_code)[T.A10S]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A12M]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A14O]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A16W]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A1CG]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A1FI]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A1VM]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A1XU]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A1YO]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A1YQ]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A20C]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A21Q]                          6.018e+10   9.92e+10
0.607      0.544   -1.34e+11    2.55e+11
C(store_code)[T.A23E]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A24S]                          -5.66e+09   9.33e+09
-0.607      0.544   -2.39e+10    1.26e+10
C(store_code)[T.A27A]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A2BQ]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A2IO]                            -0.1833      0.324
-0.565      0.572      -0.819       0.452
C(store_code)[T.A2KX]                         -2.386e+10   3.93e+10
-0.607      0.544   -1.01e+11    5.32e+10
C(store_code)[T.A2LQ]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A2NF]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A2PM]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A2RA]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
```

```
C(store_code)[T.A2WN]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A2YX]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A33J]                             0.0633      0.257
0.247      0.805      -0.440       0.566
C(store_code)[T.A357]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A360]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A364]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A39M]                          7.274e+10    1.2e+11
0.607      0.544   -1.62e+11    3.08e+11
C(store_code)[T.A3BA]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3BU]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A3DI]                         -2.386e+10   3.93e+10
-0.607      0.544   -1.01e+11    5.32e+10
C(store_code)[T.A3FS]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3FU]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3GK]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A3LU]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3PS]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3RZ]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3SS]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A3VU]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A3YY]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A408]                             0.0077      0.319
0.024      0.981      -0.617       0.633
C(store_code)[T.A409]                             0.1998      0.298
0.671      0.502      -0.383       0.783
C(store_code)[T.A412]                            -0.0483      0.333
-0.145      0.885      -0.701       0.604
C(store_code)[T.A416]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A428]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A42T]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
```

```
C(store_code)[T.A437]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A44G]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A459]                         -2.386e+10   3.93e+10
-0.607      0.544   -1.01e+11    5.32e+10
C(store_code)[T.A45W]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A45X]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A46A]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A46P]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A46R]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A470]                         -5.66e+09   9.33e+09
-0.607      0.544   -2.39e+10    1.26e+10
C(store_code)[T.A474]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A478]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A479]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A480]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A483]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A484]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A486]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A488]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A48C]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A493]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A494]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A4AP]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A4AQ]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A4AR]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A4AT]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
C(store_code)[T.A4BE]                           7.76e+09   1.28e+10
0.607      0.544   -1.73e+10    3.28e+10
```

```
C(store_code)[T.A4BY]                          6.018e+10   9.92e+10
0.607      0.544   -1.34e+11    2.55e+11
C(store_code)[T.A4CT]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A4JT]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A4KK]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A4ZA]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A505]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A517]                           4.15e+10   6.84e+10
0.607      0.544   -9.26e+10    1.76e+11
C(store_code)[T.A521]                           4.15e+10   6.84e+10
0.607      0.544   -9.26e+10    1.76e+11
C(store_code)[T.A54K]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A559]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A55E]                           4.15e+10   6.84e+10
0.607      0.544   -9.26e+10    1.76e+11
C(store_code)[T.A572]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A601]                          -5.66e+09   9.33e+09
-0.607      0.544   -2.39e+10    1.26e+10
C(store_code)[T.A624]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A651]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A657]                           4.15e+10   6.84e+10
0.607      0.544   -9.26e+10    1.76e+11
C(store_code)[T.A660]                           4.15e+10   6.84e+10
0.607      0.544   -9.26e+10    1.76e+11
C(store_code)[T.A661]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A682]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A720]                             0.1812      0.226
0.801      0.423      -0.262       0.625
C(store_code)[T.A735]                          8.137e+09   1.34e+10
0.607      0.544   -1.81e+10    3.44e+10
C(store_code)[T.A736]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.A760]                          1.639e+10    2.7e+10
0.607      0.544   -3.65e+10    6.93e+10
C(store_code)[T.E0MK]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E0SP]                         -6.302e+10   1.04e+11
-0.607      0.544   -2.67e+11    1.41e+11
```

```
C(store_code)[T.E0YT]                          4.025e+10   6.63e+10
0.607      0.544   -8.98e+10     1.7e+11
C(store_code)[T.E1HF]                         -1.597e+09   2.63e+09
-0.607      0.544   -6.76e+09    3.56e+09
C(store_code)[T.E215]                          3.566e+10   5.88e+10
0.607      0.544   -7.95e+10    1.51e+11
C(store_code)[T.E220]                          5.547e+10   9.14e+10
0.607      0.544   -1.24e+11    2.35e+11
C(store_code)[T.E228]                          4.025e+10   6.63e+10
0.607      0.544   -8.98e+10     1.7e+11
C(store_code)[T.E234]                          4.025e+10   6.63e+10
0.607      0.544   -8.98e+10     1.7e+11
C(store_code)[T.E255]                         -6.302e+10   1.04e+11
-0.607      0.544   -2.67e+11    1.41e+11
C(store_code)[T.E258]                         -6.302e+10   1.04e+11
-0.607      0.544   -2.67e+11    1.41e+11
C(store_code)[T.E260]                         -6.302e+10   1.04e+11
-0.607      0.544   -2.67e+11    1.41e+11
C(store_code)[T.E262]                         -6.302e+10   1.04e+11
-0.607      0.544   -2.67e+11    1.41e+11
C(store_code)[T.E266]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E26P]                          4.025e+10   6.63e+10
0.607      0.544   -8.98e+10     1.7e+11
C(store_code)[T.E272]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E274]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E275]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E278]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E27L]                         -6.302e+10   1.04e+11
-0.607      0.544   -2.67e+11    1.41e+11
C(store_code)[T.E281]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E283]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E289]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E296]                          -6.63e+10   1.09e+11
-0.607      0.544    -2.8e+11    1.48e+11
C(store_code)[T.E2IH]                          4.025e+10   6.63e+10
0.607      0.544   -8.98e+10     1.7e+11
C(store_code)[T.E2KL]                         -4.446e+09   7.33e+09
-0.607      0.544   -1.88e+10    9.92e+09
C(store_code)[T.E2LZ]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E2MA]                         -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
```

```
C(store_code)[T.E2TT]                          -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E310]                          -1.597e+09   2.63e+09
-0.607      0.544   -6.76e+09    3.56e+09
C(store_code)[T.E312]                          -1.597e+09   2.63e+09
-0.607      0.544   -6.76e+09    3.56e+09
C(store_code)[T.E315]                          -1.597e+09   2.63e+09
-0.607      0.544   -6.76e+09    3.56e+09
C(store_code)[T.E321]                          -1.597e+09   2.63e+09
-0.607      0.544   -6.76e+09    3.56e+09
C(store_code)[T.E322]                          -1.597e+09   2.63e+09
-0.607      0.544   -6.76e+09    3.56e+09
C(store_code)[T.E323]                           2.088e+10   3.44e+10
0.607      0.544   -4.66e+10    8.83e+10
C(store_code)[T.E329]                           2.088e+10   3.44e+10
0.607      0.544   -4.66e+10    8.83e+10
C(store_code)[T.E347]                          -4.446e+09   7.33e+09
-0.607      0.544   -1.88e+10    9.92e+09
C(store_code)[T.E348]                          -4.446e+09   7.33e+09
-0.607      0.544   -1.88e+10    9.92e+09
C(store_code)[T.E35F]                          -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E38J]                          -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.E3EP]                           4.025e+10   6.63e+10
0.607      0.544   -8.98e+10     1.7e+11
C(store_code)[T.E3HH]                          -5.749e+10   9.48e+10
-0.607      0.544   -2.43e+11    1.28e+11
C(store_code)[T.J0DT]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0FH]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0HP]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0LL]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0MZ]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0NT]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0TD]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0US]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J0VL]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J12J]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J13X]                           8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
```

```
C(store_code)[T.J146]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J147]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J152]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J154]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J155]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J159]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J15L]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J15M]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J15N]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J161]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J163]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J169]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J170]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J175]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J176]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J17T]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J185]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J187]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J188]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J18N]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J192]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J193]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J195]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J19H]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1BQ]                                  8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
```

```
C(store_code)[T.J1FM]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1GF]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1GH]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1KV]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1KW]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1LP]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J1MJ]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J202]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J400]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.J688]                                 8.853e+09   1.46e+10
0.607      0.544   -1.97e+10    3.75e+10
C(store_code)[T.L022]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L027]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L040]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L0GC]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L0MD]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L0ML]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L0PF]                                 1.772e+11   2.92e+11
0.607      0.544   -3.95e+11     7.5e+11
C(store_code)[T.L0TV]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L0YM]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L1DV]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L1HS]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L1HT]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L1LP]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L1RV]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L1ZJ]                                 2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
```

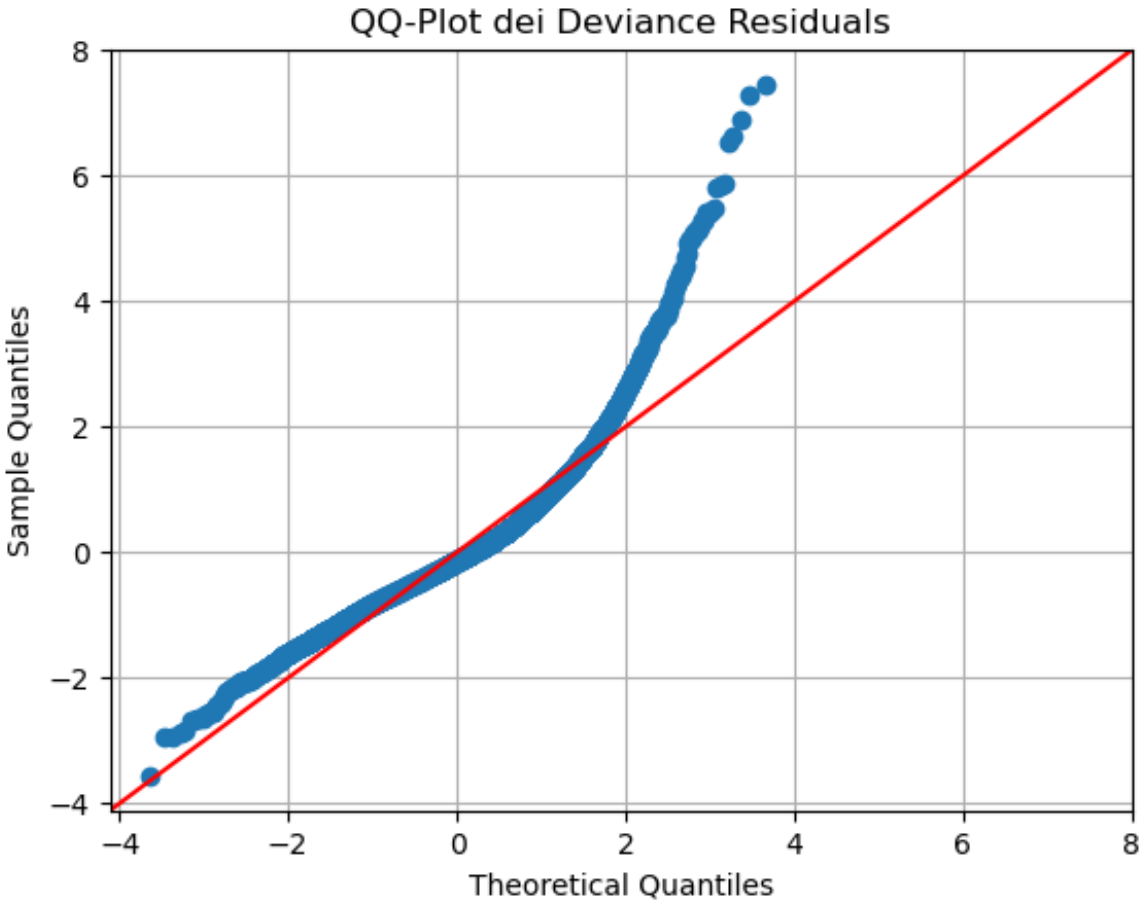
```
C(store_code)[T.L24U]                                  2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L29J]                                  2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L552]                                  2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L564]                                  2.337e+11   3.85e+11
0.607      0.544   -5.21e+11    9.89e+11
C(store_code)[T.L648]                                  2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.L763]                                  2.572e+10   4.24e+10
0.607      0.544   -5.74e+10    1.09e+11
C(store_code)[T.U054]                                  3.965e+10   6.53e+10
0.607      0.544   -8.84e+10    1.68e+11
C(store_code)[T.U056]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U061]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U064]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U066]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U068]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U072]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U078]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U080]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U084]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U086]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U089]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U095]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U097]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U099]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U0BL]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U0SK]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U0XU]                                  3.965e+10   6.53e+10
0.607      0.544   -8.84e+10    1.68e+11
C(store_code)[T.U104]                                  1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
```

```
C(store_code)[T.U105]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U10C]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U112]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U120]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U124]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U131]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U134]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U137]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U139]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1JI]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1JJ]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1JK]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1LR]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1OS]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1WO]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U1XY]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U20H]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U22P]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U23L]                          3.965e+10   6.53e+10
0.607      0.544   -8.84e+10    1.68e+11
C(store_code)[T.U581]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
C(store_code)[T.U628]                          1.239e+10   2.04e+10
0.607      0.544   -2.76e+10    5.24e+10
==============================================================================
==============================================
```

In [36]:

```
import scipy.stats as stats

resid = poisson_model.resid_deviance.copy()

# QQ plot
sm.qqplot(resid, line='45', fit=True)
```

```
plt.title("QQ-Plot dei Deviance Residuals")
plt.grid(True)
plt.show()
```

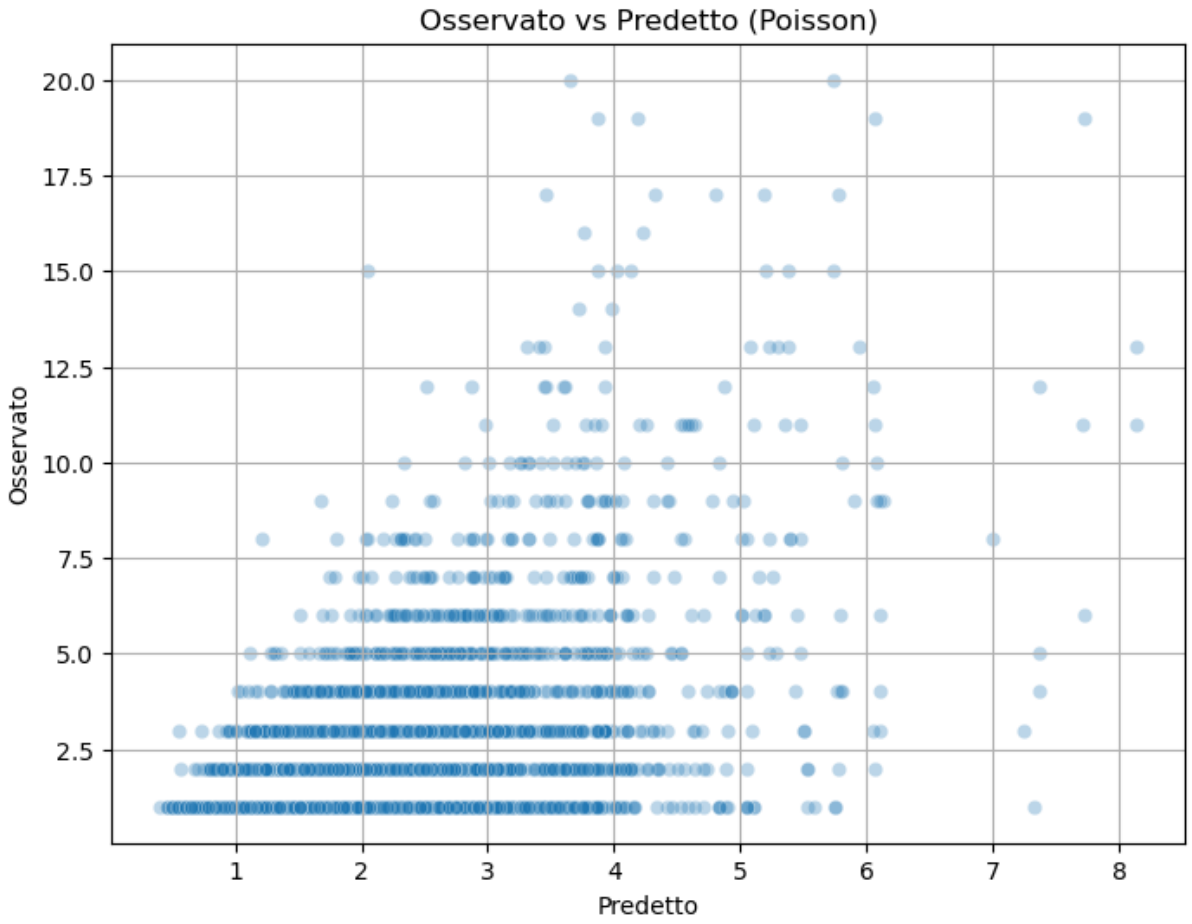


In [37]:

```
pearson_chi2 = poisson_model.pearson_chi2
df_resid = poisson_model.df_resid
dispersion = pearson_chi2 / df_resid

print("Pearson Chi2:", pearson_chi2)
print("Residual DF:", df_resid)
print("Dispersion parameter:", dispersion)
Pearson Chi2: 6042.390363640775
Residual DF: 7432
Dispersion parameter: 0.8130234612003195
```

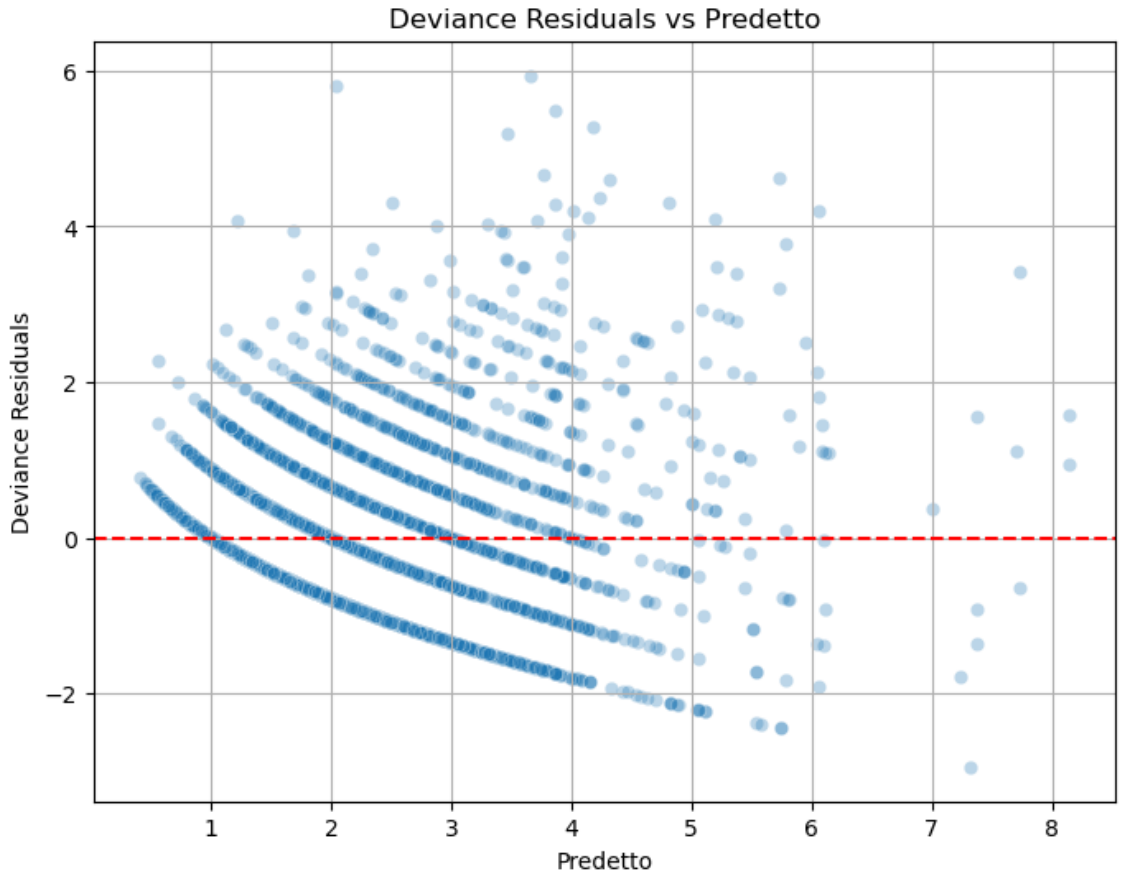
In [38]:

```
df_poisson["predicted"] = poisson_model.predict(df_poisson)

plt.figure(figsize=(8,6))
sns.scatterplot(x=df_poisson["predicted"],
y=df_poisson["sold_qty"], alpha=0.3)
plt.xlabel("Predetto")
plt.ylabel("Osservato")
plt.title("Osservato vs Predetto (Poisson)")
plt.grid(True)
plt.show()
```



In [39]:

```
plt.figure(figsize=(8,6))
sns.scatterplot(x=df_poisson["predicted"], y=resid, alpha=0.3)
plt.axhline(0, color="red", linestyle="--")
plt.xlabel("Predetto")
plt.ylabel("Deviance Residuals")
plt.title("Deviance Residuals vs Predetto")
plt.grid(True)
plt.show()
```



In [40]:

```
df_high_sales = df[df['sold_qty'] > 20].copy()
print(df_high_sales.shape)
(12, 54)
```

In [41]:

```
formula_high_sales = """sold_qty ~ C(region_desc) + C(country) +
            C(price_range) + C(macro_colour_group) +
            C(macro_merc_typology_desc) + C(dimension_group_desc)
+ C(macro_line_grouped) """
```

In [42]:

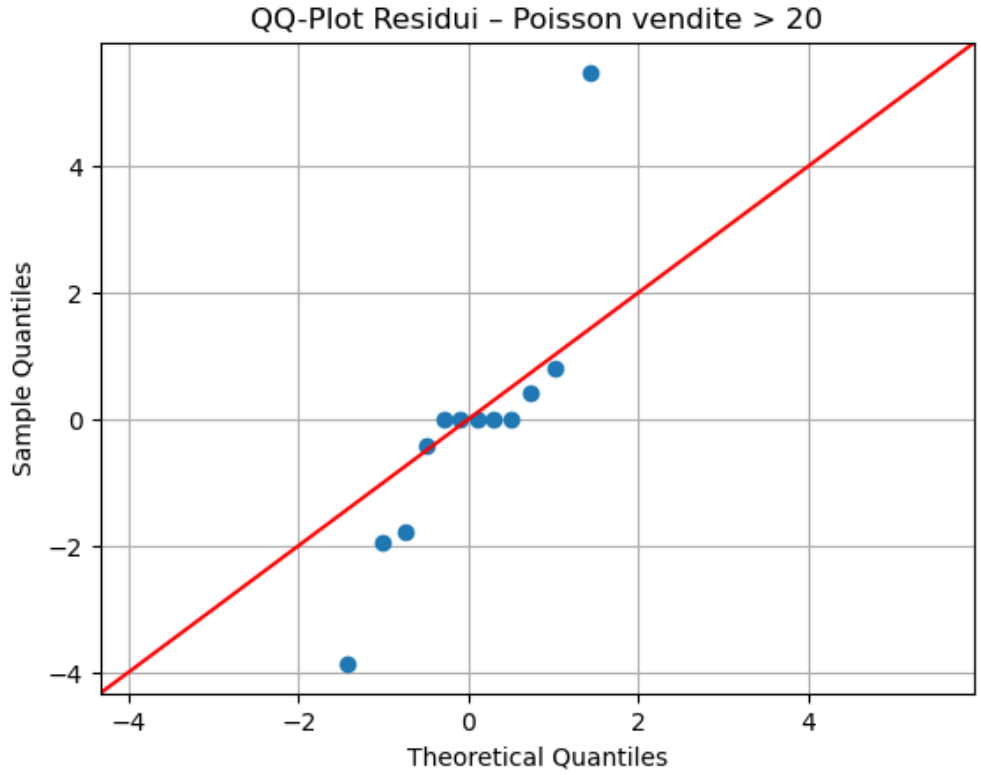
```
import statsmodels.formula.api as smf

poisson_high_sales = smf.glm(formula=formula_high_sales,
                             data=df_high_sales,
                             family=sm.families.Poisson()).fit()

print(poisson_high_sales.summary())
                 Generalized Linear Model Regression Results
==============================================================================
============
Dep. Variable:                sold_qty   No. Observations:
12
Model:                            GLM   Df Residuals:
5
Model Family:                 Poisson   Df Model:
6
```

```
Link Function:                    Log   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-58.048
Date:                Fri, 19 Sep 2025   Deviance:
52.773
Time:                        19:48:20   Pearson chi2:
57.5
No. Iterations:                     4   Pseudo R-squ. (CS):
0.9685
Covariance Type:            nonrobust
==============================================================================
                                                 coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                                     1.2331      0.047
26.477      0.000       1.142       1.324
C(region_desc)[T.Japan]                       0.4867      0.040
12.258      0.000       0.409       0.565
C(country)[T.Italy]                           0.6375      0.159
4.013      0.000       0.326       0.949
C(country)[T.Japan]                           0.4867      0.040
12.258      0.000       0.409       0.565
C(country)[T.United Kingdom]                 -0.1335      0.262
-0.510      0.610      -0.646       0.379
C(price_range)[T.CORE MEDIUM]                 0.2964      0.080
3.685      0.000       0.139       0.454
C(price_range)[T.ENTRY]                       0.3973      0.066
6.035      0.000       0.268       0.526
C(macro_colour_group)[T.GREEN]               -0.4500      0.110
-4.084      0.000      -0.666      -0.234
C(macro_colour_group)[T.NEUTRAL]              0.7464      0.070
10.590      0.000       0.608       0.885
C(macro_colour_group)[T.RED_PINK]             0.5395      0.101
5.352      0.000       0.342       0.737
C(macro_merc_typology_desc)[T.Top Handle]     0.0894      0.065
1.372      0.170      -0.038       0.217
C(macro_merc_typology_desc)[T.Tote]           0.3973      0.066
6.035      0.000       0.268       0.526
C(dimension_group_desc)[T.Small]              0.3310      0.070
4.697      0.000       0.193       0.469
C(macro_line_grouped)[T.FERRAGAMO HUG]        0.9021      0.060
15.009      0.000       0.784       1.020
C(macro_line_grouped)[T.NEW LINE AI24]       -0.0663      0.104
-0.637      0.524      -0.270       0.138
==============================================================================
==============================================
```

In [43]:

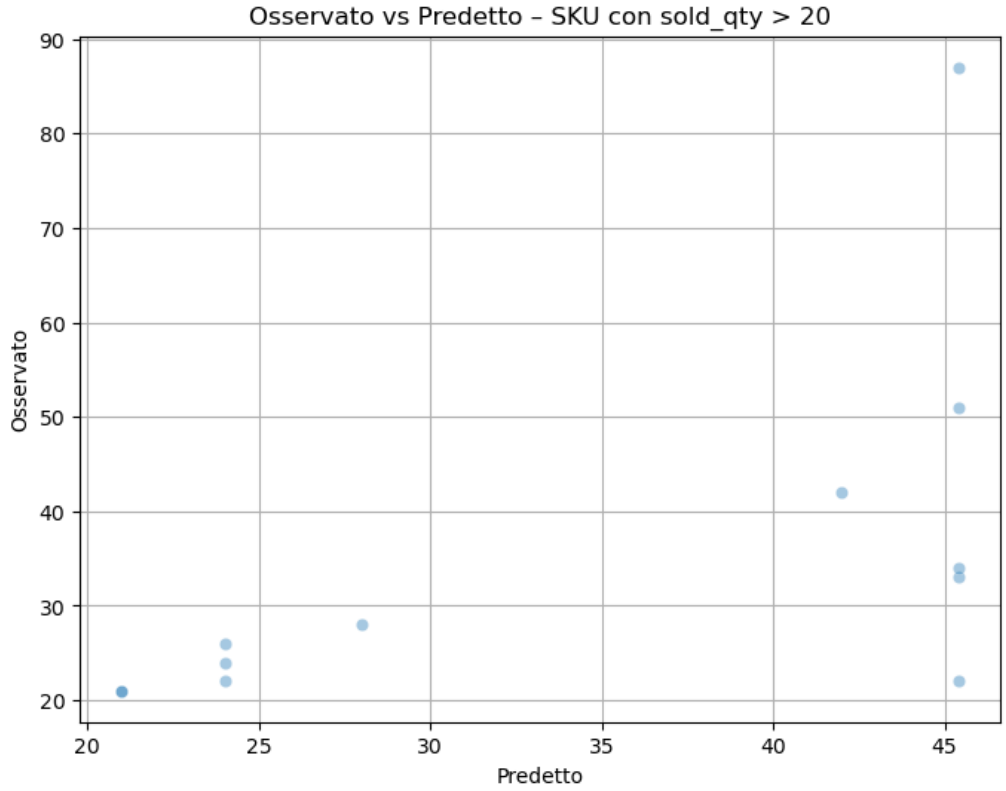
```
import statsmodels.api as sm
```

```
import matplotlib.pyplot as plt
sm.qqplot(poisson_high_sales.resid_deviance, line='45')
plt.title("QQ-Plot Residui - Poisson vendite > 20")
plt.grid(True)
plt.show()
```



In [44]:

```
plt.figure(figsize=(8,6))
sns.scatterplot(x=poisson_high_sales.fittedvalues,
                y=df_high_sales["sold_qty"],
                alpha=0.4)
plt.xlabel("Predetto")
plt.ylabel("Osservato")
plt.title("Osservato vs Predetto - SKU con sold_qty > 20")
plt.grid(True)
plt.show()
```



In [45]:

```
import joblib
```

In [46]:

```
import joblib


joblib.dump(clf, 'modello_logistico.pkl')


joblib.dump(X.columns.tolist(), 'features_logistico.pkl')
```

Out[46]:

```
['features_logistico.pkl']
```

In [47]:

```
import pickle

#  modello GLM Poisson (vendite <= 20)
with open('modello_poisson.pkl', 'wb') as f:
    pickle.dump(poisson_model, f)
```

In [48]:

```
# modello GLM Poisson (vendite > 20)
with open('modello_poisson_high_sales.pkl', 'wb') as f:
    pickle.dump(poisson_high_sales, f)
```

In [ ]:

TEST MODELLI

```
import pandas as pd
import numpy as np
import seaborn as sns
pd.set_option('display.max_columns', 500)
```

In [24]:

```
import os
print(os.getcwd())
```

```
/Users/stefanolandolfi/Desktop/FERRAGAMO_PYTHON/01_python
```

In [25]:

```
# Logistic regression
import joblib
clf = joblib.load('modello_logistico.pkl')
features_logistico = joblib.load('features_logistico.pkl')

# Poisson
import pickle
with open('modello_poisson.pkl', 'rb') as f:
    poisson_model = pickle.load(f)

with open('modello_poisson_high_sales.pkl', 'rb') as f:
    poisson_high_sales = pickle.load(f)
```

In [26]:

```
df_test =
pd.read_excel("/Users/stefanolandolfi/Desktop/test_data_50_store.x
lsx", sheet_name="Sheet1")
```

In [27]:

```
df_test.head()
```

Out[27]:

	region_desc	country	price_range	macro_colour_group	macro_merc_typology_desc	dimension_group_desc	macro_line_grouped	store_code
0	China	China	ENTRY	LILAC	Minibag	Mini	FLORENCE	A0AW
1	Latin America	Chile	ENTRY	LILAC	Minibag	Medium	FERRAGAMO HUG	L0PF
2	SEAP	Australia	CORE MEDIUM	NEUTRAL	Clutch	Small	FERRAGAMO ARCHIVE	A0AL
3	Japan	Japan	ENTRY	NEUTRAL	Top Handle	Medium	RAINBOW MATELASSE'	J0DT

	region_desc	country	price_range	macro_colour_group	macro_merc_typology_desc	dimension_group_desc	macro_line_grouped	store_code
4	Europe	Belgium	TOP	YELLOW_GOLD	Shoulder Bag	Small	STAR	E220

In [28]:

```
X_test = pd.get_dummies(df_test, drop_first=True)


for col in features_logistico:
    if col not in X_test.columns:
        X_test[col] = 0

X_test = X_test[features_logistico]
```

In [29]:

```
df_test['prob_vendita'] = clf.predict_proba(X_test)[:, 1]
```

In [33]:

```
import statsmodels.formula.api as smf


formula = """sold_qty ~ C(region_desc) + C(country) +
             C(price_range) + C(macro_colour_group) +
             C(macro_merc_typology_desc) + C(dimension_group_desc)
+ C(macro_line_grouped) + C(store_code)"""


df_test['vendite_stimate'] = poisson_model.predict(df_test)
```

In [34]:

```
df_test.head(10)
```

Out[34]:

	region_desc	country	price_range	macro_colour_group	macro_merc_typology_desc	dimension_group_desc	macro_line_grouped	store_code	prob_vendita	vendite_stimate
(	China	China	ENTRY	LILAC	Minibag	Mini	FLORENCE	A0AW	0.013391	1.097142
1	Latin America	Chile	ENTRY	LILAC	Minibag	Medium	FERRAGAMO HUG	L0PF	0.152561	1.457504
2	SEAP	Australia	CORE MEDIUM	NEUTRAL	Clutch	Small	FERRAGAMO ARCHIVE	A0AL	0.993419	0.698965
3	Japan	Japan	ENTRY	NEUTRAL	Top Handle	Medium	RAINBOW MATELASSE'	J0DT	0.002483	1.244824

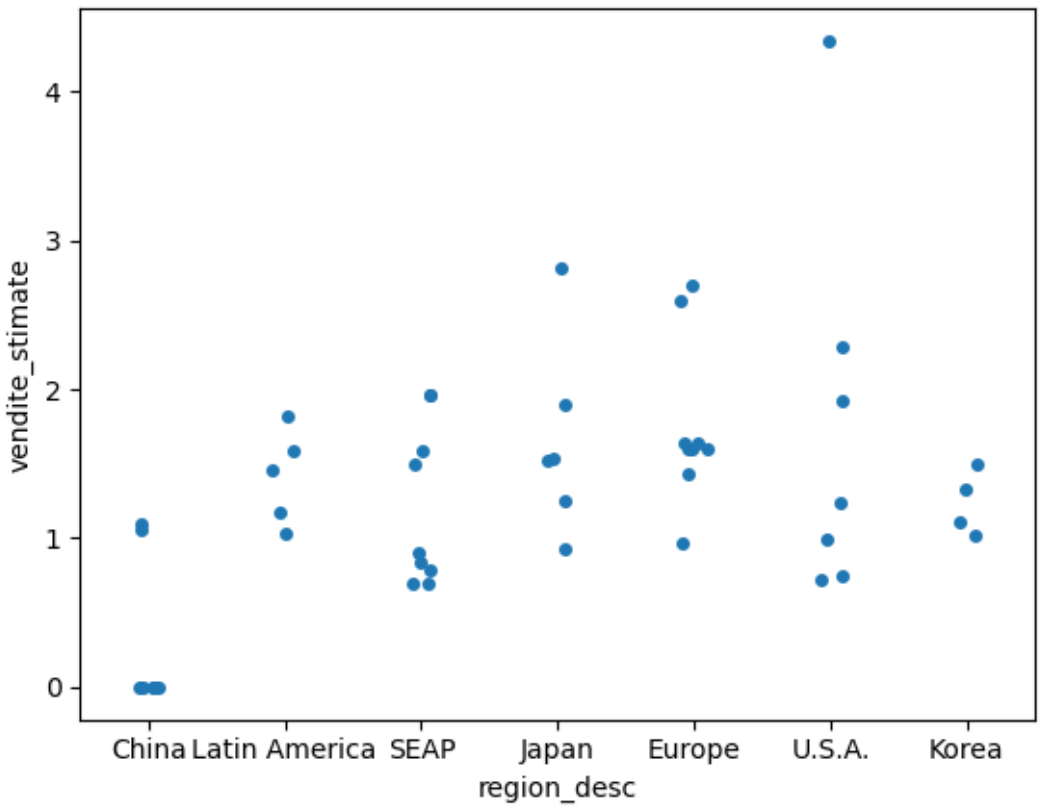
	region_desc	country	price_range	macro_colour_group	macro_merc_typology_desc	dimension_group_desc	macro_line_grouped	store_code	prob_vendita	vendite_stimate
4	Europe	Belgium	TOP	YELLOW_GOLD	Shoulder Bag	Small	STAR	E220	0.999924	1.598776
5	SEAP	Singapore	HIGH	BROWN	Minibag	Mini	FERRAGAMO CREATION	A24S	0.999972	1.496969
6	Japan	Japan	TOP	RED_PINK	Hobo	Small	FLORENCE	J0DT	0.999958	1.539312
7	Japan	Japan	HIGH	RED_PINK	Clutch	Small	NEW LINE AI24	J0DT	0.999992	1.523702
8	Japan	Japan	TOP	TURQUOISE	Minibag	Small	FERRAGAMO CREATION	J0DT	0.999996	0.924905
9	Europe	France	HIGH	GREEN	Tote	Mini	WANDA	E0YT	0.999936	1.642895

In [35]:

```
sns.stripplot(data=df_test, x='region_desc', y='vendite_stimate')
```

Out[35]:

```
<Axes: xlabel='region_desc', ylabel='vendite_stimate'>
```



utile come indicatore di performance per capire quali sono gli sku che performano meglio, ma troppi limiti per la quantità da vendere, servirebbe una raccolta dati che sia consona al lavoro. quale sku avrà performance ottimali a seconda dello store, con la clusterizzazione gia testata dall'azinda si sa quanti sku possono andare in un determinato store. Quindi si scelgono i migliori sku per ogni store a seconda delle linee e poi si restringe il campo fino a proporre n sku ai buyer

TEST MODELLI REAL DATA

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
pd.set_option('display.max_columns', 500)
```

In [2]:

```
import os
print(os.getcwd())
/Users/stefanolandolfi/Desktop/FERRAGAMO_PYTHON/01_python
```

In [3]:

```
# Logistic regression
import joblib
clf = joblib.load('modello_logistico.pkl')
features_logistico = joblib.load('features_logistico.pkl')

# Poisson
import pickle
with open('modello_poisson.pkl', 'rb') as f:
    poisson_model = pickle.load(f)

with open('modello_poisson_high_sales.pkl', 'rb') as f:
    poisson_high_sales = pickle.load(f)
```

In [4]:

```
df_test =
pd.read_excel("/Users/stefanolandolfi/Desktop/Borse_per_store_ferr
agamo.xlsx", sheet_name="Sheet1")
```

In [5]:

```
df_test.head()
```

Out[5]:

	country	store_code	region_desc	dimension_group_desc	macro_line_grouped	colour_group	price_range	macro_merc_typology_desc	sku_code
0	Canada	U054	U.S.A.	Medium	THE STUDIO	WHITE	HIGH	Tote	786555
1	Canada	U054	U.S.A.	Medium	THE STUDIO	BROWN	HIGH	Tote	786556

	country	store_code	region_desc	dimension_group_desc	macro_line_grouped	colour_group	price_range	macro_merc_typology_desc	sku_code
2	Canada	U054	U.S.A.	Large	THE STUDIO	BLACK	HIGH	Tote	786557
3	Canada	U054	U.S.A.	Large	THE STUDIO	MULTI COLOUR	TOP	Tote	786558
4	Canada	U054	U.S.A.	Large	FERRAGAMO HUG	DARK BROWN	HIGH	Top Handle	786559

In [6]:

```
colour_mapping = {
    'BLACK': 'NEUTRAL', 'WHITE': 'NEUTRAL', 'BEIGE': 'NEUTRAL',
'TAUPE': 'NEUTRAL',
    'GREY': 'NEUTRAL', 'LIGHT GREY': 'NEUTRAL', 'DARK GREY':
'NEUTRAL', 'CAMEL': 'NEUTRAL',
    'BONE': 'NEUTRAL', 'SILVER': 'NEUTRAL', 'TRANSPARENT':
'NEUTRAL', 'NOT DEFINED': 'NEUTRAL',

    'DARK BROWN': 'BROWN', 'MEDIUM BROWN': 'BROWN', 'TAN':
'BROWN', 'RUST': 'BROWN', 'ORANGE': 'BROWN', 'BROWN': 'BROWN',

    'RED': 'RED_PINK', 'BORDEAUX': 'RED_PINK', 'LIGHT PINK':
'RED_PINK', 'DARK PINK': 'RED_PINK', 'PURPLE': 'RED_PINK',

    'BLUE': 'BLUE', 'BLUETTE': 'BLUE', 'LIGHT BLUE': 'BLUE',

    'DK OLIVE GREEN': 'GREEN', 'LT OLIVE GREEN': 'GREEN', 'DK
BRILL.GREEN': 'GREEN',
    'LT BRILL.GREEN': 'GREEN', 'DARK AQUA GREEN': 'GREEN', 'LIGHT
AQUA GREEN': 'GREEN',

    'LIGHT YELLOW': 'YELLOW_GOLD', 'DARK YELLOW': 'YELLOW_GOLD',
'GOLD': 'YELLOW_GOLD',

    'LILAC': 'LILAC', 'TURQUOISE': 'TURQUOISE',
}

df_test['macro_colour_group'] =
df_test['colour_group'].map(colour_mapping)

print(df_test[['colour_group', 'macro_colour_group']].head(10))
```

```
   colour_group macro_colour_group
0         WHITE            NEUTRAL
1         BROWN              BROWN
2         BLACK            NEUTRAL
3   MULTICOLOUR                NaN
4    DARK BROWN              BROWN
5  MEDIUM BROWN              BROWN
```

```
6  LIGHT YELLOW        YELLOW_GOLD
7         BLACK            NEUTRAL
8    DARK BROWN              BROWN
9  LIGHT YELLOW        YELLOW_GOLD
```

In [7]:

```
X_test = pd.get_dummies(df_test, drop_first=True)


for col in features_logistico:
    if col not in X_test.columns:
        X_test[col] = 0

X_test = X_test[features_logistico]
```

In [8]:

```
df_test['prob_vendita'] = clf.predict_proba(X_test)[:, 1]
```

In [14]:

```
import statsmodels.formula.api as smf

formula = """sold_qty ~ C(region_desc) + C(country) +
             C(price_range) + C(macro_colour_group) +
             C(macro_merc_typology_desc) + C(dimension_group_desc)
+ C(macro_line_grouped) + C(store_code)"""


df_test['vendite_stimate'] = poisson_model.predict(df_test)
```

In [16]:

```
df_test.head(10)
```

Out[16]:

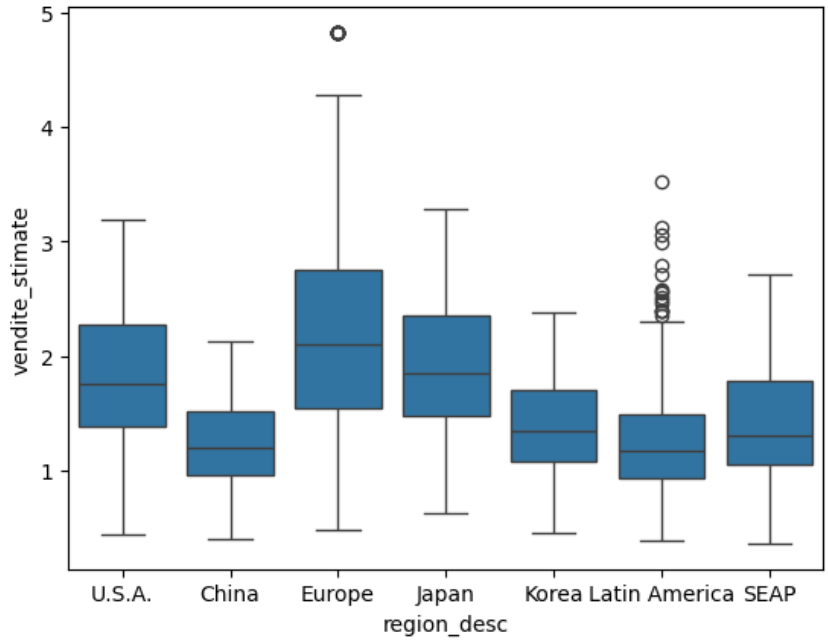
	country	store_code	region_desc	dimension_group_desc	macro_line_grouped	colour_group	price_range	macro_merc_typology_desc	sku_code	macro_colour_group	prob_vendita	vendite_stimate
0	Canada	U054	U.S.A.	Medium	THE STUDIO	WHITE	HIGH	Tote	7865555	NEUTRAL	0.9993341	1.274709
1	Canada	U054	U.S.A.	Medium	THE STUDIO	BROWN	HIGH	Tote	7865556	BROWN	0.9995542	1.609661
2	Canada	U054	U.S.A.	Large	THE STUDIO	BLACK	HIGH	Tote	7865557	NEUTRAL	0.9995542	0.966484
3	Canada	U054	U.S.A.	Large	THE STUDIO	MULTICOLOUR	TOP	Tote	7865558	NaN	0.9990031	NaN
4	Canada	U054	U.S.A.	Large	FERRAGAMO HUG	DARK BROWN	HIGH	Top Handle	7865559	BROWN	0.9997840	1.067273

	country	store_code	region_desc	dimension_group_desc	macro_line_grouped	colour_group	price_range	macro_merc_typology_desc	sku_code	macro_colour_group	prob_vendita	vendite_stimate
4	Canada	U054	U.S.A.	Medium	FERRAGAMO HUG	MEDIUM BROWN	HIGH	Top Handle	786560	BROWN	0.9968892	1.407641
6	Canada	U054	U.S.A.	Mini	FERRAGAMO HUG	LIGHT YELLOW	CORE HIGH	Minibag	786561	YELLOW_GOLD	0.9997277	1.692846
7	Canada	U054	U.S.A.	Medium	FERRAGAMO HUG	BLACK	HIGH	Shoulder Bag	786562	NEUTRAL	0.9985338	1.567158
8	Canada	U054	U.S.A.	Large	FERRAGAMO HUG	DARK BROWN	HIGH	Top Handle	786563	BROWN	0.9978440	1.067273
9	Canada	U054	U.S.A.	Mini	FERRAGAMO HUG	LIGHT YELLOW	CORE MEDIUM	Minibag	786564	YELLOW_GOLD	0.9999915	1.669859

Qui è presente uno dei GAP principali di questa ricerca. La raccolta dati per far si che un modello possa funzionare deve essere resa uniforme per il modello. il multicolour e altre caratteristiche sono classificate come NaN essendo delle nuove caratteristiche. Ferragamo potrebbe classificarle per allenare in modo adeguato il modello

In [17]:
```
sns.boxplot(data=df_test, x='region_desc', y='vendite_stimate')
```

Out[17]:
```
<Axes: xlabel='region_desc', ylabel='vendite_stimate'>
```



In [27]:

```
top_sku_per_store = (
    df_test
    .groupby(['region_desc', 'store_code', 'sku_code'],
as_index=False)
    .agg({'vendite_stimate': 'sum'})
    .sort_values(['region_desc', 'store_code', 'vendite_stimate'],
ascending=[True, True, False])
)

top15_per_store = (
    top_sku_per_store
    .groupby(['region_desc', 'store_code'])
    .head(15)
)
```

In [28]:

```
top15_per_store
```

Out[28]:

	region_desc	store_code	sku_code	vendite_stimate
15	China	A01N	786570	2.127900
34	China	A01N	786589	1.886862
32	China	A01N	786587	1.845652
31	China	A01N	786586	1.810551
16	China	A01N	786571	1.685313
...	...	...	...	...
12579	U.S.A.	U628	786556	2.223884
12585	U.S.A.	U628	786562	2.165220
12606	U.S.A.	U628	786583	2.165220
12583	U.S.A.	U628	786560	1.944804
12600	U.S.A.	U628	786577	1.836701

4980 rows × 4 columns

In [25]:

```
top_sku_per_store = (
    df_test
    .groupby(['region_desc', 'store_code', 'sku_code'],
as_index=False)
    .agg({'prob_vendita': 'sum'})
    .sort_values(['region_desc', 'store_code', 'prob_vendita'],
ascending=[True, True, False])
)

top15_per_store_prob = (
    top_sku_per_store
    .groupby(['region_desc', 'store_code'])
    .head(15)
)
```

In [26]:

```
top15_per_store_prob
```

Out[26]:

	region_desc	store_code	sku_code	prob_vendita
10	China	A01N	786565	0.999970
29	China	A01N	786584	0.999849
27	China	A01N	786582	0.999829
9	China	A01N	786564	0.999814
15	China	A01N	786570	0.999734
...	...	...	...	...
12590	U.S.A.	U628	786567	0.999752
12591	U.S.A.	U628	786568	0.999752
12589	U.S.A.	U628	786566	0.999740
12612	U.S.A.	U628	786589	0.999740
12584	U.S.A.	U628	786561	0.999682

4980 rows × 4 columns

Il modello risponde bene, gli sku in questione sono quelli che per Ferragamo ora stanno registrando un ottimo risultato e sono stati acquistati dai buyer. Prendendo un esempio un sku(codice test non dato da Ferragamo) è la hug modello di punta di Ferragamo della stagione 24/25

In [32]:

```
top_stores = df_test[df_test['store_code'].isin(['A01N', 'A516'])]


top5_sku_stores = (
    top_stores
    .groupby(['store_code', 'sku_code'], as_index=False)
    .agg({'vendite_stimate': 'sum'})
```

```
    .sort_values(['store_code', 'vendite_stimate'],
ascending=[True, False])
    .groupby('store_code')
    .head(5)
)
```

In [33]:

```
top5_sku_stores
```

Out[33]:

	store_code	sku_code	vendite_stimate
15	A01N	786570	2.127900
34	A01N	786589	1.886862
32	A01N	786587	1.845652
31	A01N	786586	1.810551
16	A01N	786571	1.685313
54	A516	786571	2.636579
73	A516	786590	2.337877
71	A516	786588	2.286789
70	A516	786587	2.243379
55	A516	786572	2.088254

Come si può notare qui sono stati presi in esame due store diversi, le vendite non coincidono, un grande risultato poiché conferma che ogni mercato opera secondo le sue leggi e preferenze. Ma si può notare come alcuni sku sono presenti in entrambe le top5 per performance predette, ottimo per stabilire un Global core assortment ed accontentare i buyer

In questo caso sono stati presi China e Seap in esame.