INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Through Costumers' Thirst - A Sales Forecasting of Spirit Drinks

João Francisco Teixeira Rodrigues

Master in Data Science

Supervisor:
Dr. Diana Mendes, Associate Professor
ISCTE-IUL

Supervisor:
Eng. Rita Pina, Company of Study

October, 2024

BUSINESS
SCHOOL

TECNOLOGIAS
E ARQUITETURA

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

Through Costumers' Thirst - A Sales Forecasting of Spirit Drinks

João Francisco Teixeira Rodrigues

Master in Data Science

Supervisor:
Dr. Diana Mendes, Associate Professor
ISCTE-IUL

Supervisor:
Eng. Rita Pina, Company of Study

October, 2024

I dedicate this study first to my main supervisor, Professor Diana, as the person who saw it grew and helped make it possible.

I dedicate this study to my close friends: Bárbara Ferreira, Carlos Faria, Duarte Marques, Jhonata Oliveira, Pedro Silva and Raquel Elias, that supported me on this journey in a close way.

I dedicate this study to my close family - my mother, father, sister, brother-in-law and my sweet goddaughter.

For the general readers, fill your thirst of knowledge as it is the only thing I can pour to you.

Acknowledgment

First or all, I express my sincerest gratitude to both my supervisors: Professor Diana Mendes and Eng. Rita Pina, for the support and guidance through the project. A special thanks to Professor Diana as the person who guided through all the process and with a special effort during this last month of delivery - October, and I also take the chance to state how good of a teacher and a person she is.

I thank Bacardi for proposing this intriguing real-world project to the academic community and fostering a connection with the university.

I also leave here a thanks to Gonçalo Almeida, which had a similar project, and with whom I traded ideas for the development, establishing a mutual help.

Moving to the persons that were not involved on the project, but gave some handouts, makes the days better, encourages and motivates with an unconditional support, I want to express my feelings of passion and luckiness to have this special group of friends from the Master, that I adore: Bárbara Ferreira, Carlos Faria, Duarte Marques, Jhonata Oliveira and Pedro Silva. And I also want to give a special mention to Raquel Elias, the adopted friend on the group, the group that is so cool (*haha*), that the day I am writing (last day), it is Halloween and we are going to party, after 1 hour of sleep will be interesting.

I want to thanks my mother for the remote support and my beloved 1 year old goddaughter to annoy my sister. Also, of course a thanks to my father to ask me if why it was not done yet, my sister and my brother-in-law, the closest family.

Last but not least, I want to thank myself, not in an egocentric way, but because I am proud of what I was able to achieve, after a very heavy month of my birthday, but it payed off those nights out of sleep.

Resumo

Prever antecipadamente as necessidades de *stock* para um determinado período é uma ferramenta essencial para garantir a satisfação dos clientes e evitar perdas. Esta dissertação explora e aborda um problema de previsão para a Bacardi, uma das principais empresas de bebidas espirituosas, com o objetivo de prever as vendas de vários produtos e identificar fatores chave que as influenciam, de forma a otimizar a gestão de estoque. Neste contexto, foram analisados dados de vendas dos cinco produtos mais vendidos no canal off-trade, aplicando-se modelos de previsão, incluindo o *Seasonal Auto-Regressive Integrated Moving Average* (SARIMA), Seasonal Auto-Regressive Integrated Moving Average with Exogenous factors (SARIMAX) e o *Extreme Gradient Boosting Machines* (XGBM), incorporando variáveis exógenas relevantes. Os resultados indicam que os modelos SARIMA ou SARIMAX são os mais eficazes, com o PIB, PSI, ICC e cobertura de nuvens emergindo como variáveis exógenas significativas, quando aplicável.

Abstract

Accurately forecasting stock requirements for a given period is crucial to ensuring customer satisfaction and minimizing losses. This dissertation explores and addresses a forecasting problem for Bacardi, a leading spirit drinks company, aiming to predict sales across multiple products and identify key factors influencing them to optimize stock management. In this context, sales data for the five best-selling products in the off-trade channel were analyzed, with forecasting models applied, including Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Seasonal Auto-Regressive Integrated Moving Average with Exogenous factors (SARIMAX), and Extreme Gradient Boosting Machines (XGBM), incorporating relevant exogenous variables. The results indicate that either SARIMA or SARIMAX models perform best, with GDP, PSI, ICC, and cloud cover emerging as significant exogenous variables when applicable.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Bacardi, a producer of spirits drinks, aims to minimize the products' stock on hold, as the products have a limited shelf life, and it is not desirable to incur unnecessary production costs. To achieve this goal, it is essential to know the expected sales amount beforehand, which is where sales forecasting comes into play.

Sales forecasting plays a crucial role in various managerial decisions. According to S. Wang et al. (2010), sales forecasting are "essential inputs to many managerial decisions, such as pricing, store space allocation, listing/delisting, ordering and inventory management for an item. Forecasts also provide the basis for distribution and replenishment plans". Additionally, accurate sales forecasts can "lead to improved customer satisfaction, reduced waste, increased sales revenue and more effective and efficient distribution".

Given the wide range of products offered by the brand, this study focuses, for forecasting purposes, on the top 5 best-selling SKUs by volume, over 3-months, 6-months and 12-months periods. However, an exploratory data analysis is also conducted across the entire channel.

To forecast sales, it was used SARIMA, SARIMAX and XGBMs models, having obtained better results with SARIMA and SARIMAX, and when exogenous variables had effect, they were: GDP, PSI, ICC or cloud cover.

This chapter provides an overview of the problem, discusses the generic approach to solving it, outlines the difficulties encountered, and identifies the research gap.

1.1. Contextualization

The company is a leading multinational producer in the spirits industry, renowned for its extensive market presence and wide-ranging portfolio. Its product lineup spans a variety of categories, including whiskey, vodka, rum, gin, aperitifs, vermouths, sparkling wine, and specialty wines, all marketed under a number of prominent brands. For this study, the company provided data for a single brand that includes products from the following categories: aperitifs, ready-to-serve, sparkling wine, special wines, and ready-to-drink beverages. Catering to diverse consumer preferences, the company's offerings range from casual drinks to premium selections. With a commitment to quality and innovation, the company plays a significant role in the global spirits market and strives to meet diverse consumer preferences.

Operating exclusively as a Business To Business (B2B) entity, the organization segments its sales into off-trade and on-trade channels. Off-trade includes markets such as minimarkets and supermarkets, which account for the minority of sales. On-trade encompasses consumption establishments like cafes, restaurants, bars, and hotels, representing the majority of sales.

This separation is essential for business logistics, as each channel has distinct demand patterns and operational requirements, making it necessary to make predictions by segment and by product. A product is defined by the combination of its bottle code (name) and bottle size (volume in centiliters). The scope of this study is restricted to the off-trade market.

At the beginning of the chapter, the benefits of using sales forecasting, especially for stock control, were highlighted. Despite these benefits, the company currently manages its stock using a more traditional and antiquated method, where representatives gather to review past data and, based on their experience, attempt to predict the upcoming year. This approach is prone to errors and is time-consuming, therefore, this dissertation aims to develop models that can outperform this traditional method and streamline the forecasting process.

Besides forecasting, one of the objectives is to discover relationships between sales and external factors, both social and economic. Understanding these relationships can provide valuable insights into market trends and consumer behavior, enabling the company to make more informed strategic decisions. Additionally, this knowledge can help the company anticipate and adapt to changes in the market, ultimately leading to improved sales performance and competitive advantage.

1.2. Research and Development

As previously mentioned, the goal is to forecast each product's sales volume on the off-trade channel. Although this is the main and final goal, there is an intermediate point to consider, mentioned in the last paragraph, which recalls and rephrases as a question: Which external factors affect sales and how? In this section, is described in a high-level view, the processes to achieve the answers, and the research to support it.

First, the data disposed of represents monthly sales in volume, and given that it is temporal data, it is most logical to interpret it as a time series problem and take advantage of temporal dependencies. The possible approaches to dealing with such a task can originate from statistical and machine learning methods, presented in the literature review in section 2.2.

Forecasting can be done exclusively with historical observation of time series data, but it won't be so accurate and may fail to capture some patterns. External factors may affect a variable's prediction - in this case, the volume of sales. Social and economic factors weigh consumer behavior: in the presence of an economic depression, people will spend less; a rise in unemployment might decay the buyers that got unemployed, and so, by joining these factors to the forecast, it's possible to have explanatory variables. There are various variables that can be used; some might automatically pop into our heads, like unemployment or GDP, and brainstorming is, indeed, one of the sources, other valuable fonts are articles from related studies and IWSR reports given by the company, for the years of 2022 and 2024. IWSR is a company that analyses the drinks market. From the IWRS report, it is possible to infer factors affecting different types of drinks and trends.

1.3. Difficulties and Research Gap

The goal and questions to ask are settled, and the process to solve them was also described in the previous section 1.2. Now, what are the rocks on the way, and what's the innovation on a forecasting problem?

In terms of the market, it's being explored with many intermediaries before it reaches the final consumer. Once the company of study sells to another, this one could be a retailer and sell to another company, which might or might not sell to the final customer. These intermediations lead to operations where the company of study has no direct influence or knowledge to control, and which may affect its sales. In the meeting with the company team, one of the mentioned aspects on this subject was the occurrence of a client company making a discount on a competitor's drink and how other competitors of such a company might do the same and reduce the sales of the producer company.

A literature review reveals a lack of studies on sales forecasting for spirits and beverages in general. Additionally, there is limited research on the influence of external factors on sales within this industry. Consequently, this dissertation aims to contribute to the literature by addressing these gaps and providing insights into how external factors influence sales in the spirits industry.

1.4. Dissertation Structure

The present dissertation is structured as follows:

(1) **Introduction** - Identifies the problem and goals, contextualizes the problem and its origin, and examines the difficulties encountered along with the research gap.

(2) **Literature Review** - Analyzes previous research and theoretical frameworks relevant to the study, synthesizing key findings and identifying gaps in the existing knowledge the dissertation will address.

(3) **Methodology** - Describes the approach to tackling the problem.

(4) **Exploratory Data Analysis** - Examines the data through descriptive statistical analysis and visualizations. It also includes a detailed description of the received data and the transformations applied to prepare the data for analysis.

(5) **Modeling** - Details the predictive models employed, the results obtained, and introduces the baseline models used as a starting point for comparison.

(6) **Conclusions** - Summarizes insights gained from the various models applied, highlights the key variables influencing sales, and discusses potential future work and improvements.

CHAPTER 2

Literature Review

In alignment with the objectives established for this dissertation, the literature review focuses on two primary areas: sales forecasting methods and the integration of external factors in similar contexts, specifically within the beverage industry.

Before delving into the research findings, the methodology employed for this literature review is outlined in detail. This includes the criteria for selecting relevant studies, the databases consulted, and the approach taken to synthesize the information gathered.

2.1. Research Methodology

The research started with a focus on the general world of sales forecasting, later adding the concept of demand forecast, due to their closeness. The goal with those two keywords of search was, not limited, but specially, to gather common techniques used; sectors of action covered and how they could relate to the project's topic. In this phase, for both terms, the sources were first without a focus on one specific, using diverse, for example: Google Scholar, Scopus, and Research Gate. Over time, the process converged to use a single one, being chosen Scopus.

During the process, naturally occurred a lot of articles, which led to not reading all the papers to their fullest, but filtering them, according to the title and abstract. Articles that seemed relevant would be pointed in a table which included, for each article, mainly:

- Title
- Year received
- URL for further access
- Methods used
- "Order" which is a category to prioritize its study, where 0 has the most priority, and higher labels, less
- Sector (ex. retail or pharmaceutical)
- Category that describes the area of study it relates (ex. Sales Forecasting, Demand Forecasting, Macroeconomics)

Below, in Table 1, is represented part of the table.

Name	Received Year	Order	Statistics/ ML	Statistical Models	ML Models	URL
Intermittent demand forecasting and stock control: An empirical study	2012	0	Machine Learning	Moving Average (MA) Single Exponential Smoothing (SES) Croston's method (CRO) Synthetos-Boylan Approximation (SBA)		[url]
Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion	2010	2	Machine Learning Statistics	ARIMAX ETSX	Support Vector Regression	[url]

TABLE 1. Representation of table used to organize articles of study. It was used main columns and not disposed the urls for display purposes.

After covering the forecasting section, the external factors are divided into three areas: weather, holidays and economic indicators. Holidays did not get too much attention, given it's not possible to analyze it rigorously, as described ahead, on section 2.4, the purpose here is simply to highlight their relevance on this kind of problems. In this part of the study on external factors, a different approach was taken by shifting the focus to the retail sector, though research still covered a broad range of areas. This shift was made because the study analyzes the off-trade channel, establishing a connection between retail and the off-trade channel, as retail represents the buyer.

2.2. Forecasting Methods

Sales forecasting has been a problem present in the literature, for some time now, and given the nature of the data, it is treated like a time series problem. It is to mention that during the research, was also searched demand forecast, due to its similarity with sales forecasting.

Initially, prominent statistical models such as Auto-Regressive (AR) and Moving Average (MA) models were often combined to form ARMA or ARIMA models (Babai et al., 2013; S. Wang et al., 2010). These models can also incorporate a seasonal factor, resulting in the SARIMA model. In Pongdatu and Putra (2018), the authors compare this variant with Holt-Winters' exponential smoothing, demonstrating better results with SARIMA.

Single Exponential Smoothing (SES) methods also gained attention, which Gustriansyah et al. (2019) describes as "more suitable for predicting fluctuations things randomly (irregularly)". Croston (1972) further extended exponential smoothing, developing what later became known as Croston's method. Willemain et al. (1994) concluded that "Croston's method is robustly superior to exponential smoothing and could provide tangible benefits to manufacturers forecasting intermittent demand."

However, despite the advantages of Croston's method, Syntetos and Boylan (2001) identified a bias in the approach. To address this, Syntetos and Boylan (2005) proposed what is now known as the Syntetos-Boylan Approximation (SBA), which has been adopted by other researchers as a solution to mitigate this bias.

With the advance of computation, machine learning started a dominant process and became popular to use neural networks, starting with the simplest (Alon et al., 2001; Chawla et al., 2019; Thiesing & Vornberger, 1997). Particularly, on Alon et al. (2001) it was done a comparison between Artificial Neural Networks (ANNs) with other statistic methods, from where, on average, the results were better for neural networks, essentially for its ability to capture non linear trend and seasonal patterns.

Bayesian Neural Networks (BNNs) were applied by Gaur et al. (2015) to compare their performance against K-Nearest Neighbors (KNN), with BNNs demonstrating superior results. In the context of fashion retail, Loureiro et al. (2018) utilized Deep Neural Networks (DNNs) and compared them with traditional machine learning algorithms. Although DNNs performed well, they did not surpass some models like Random Forest. Furthermore, Kuo (2001) and Kuo et al. (2016) explored Fuzzy Neural Networks (FNNs), while Krichene et al. (2017) focused on Recurrent Neural Networks (RNNs), and Bandara et al. (2019) applied Long Short-Term Memory (LSTM) networks in the context of e-commerce.

Various studies have compared statistical methods with machine learning approaches. For instance, Spiliotis et al. (2022) specifically focuses on this comparison, finding that machine learning algorithms outperform statistical methods in terms of accuracy and bias. Weng et al. (2020) notes that while statistical methods are simple and fast to compute, selecting the right method is "an uneasy task," requiring "some expert knowledge." Furthermore, they state that "in terms of performance, they do not usually lead to very promising results" and that "compared to deep learning methods, statistical models' performance is usually worse."

Despite the performance advantages of deep learning, Weng et al. (2020) also highlights a major weakness: interpretability. While deep learning models often lack transparency, statistical models offer clearer insights into how their results are obtained. This trade-off has led to the rise of hybrid models, which aim to combine the strengths of both approaches—offering both high accuracy and interpretability. Additionally, hybrid models capture both linear and non-linear components in data, overcoming the limitations of statistical models, which are typically restricted to modeling linear relationships.

Several examples of hybrid models exist in both general time series forecasting and sales forecasting. Sadaei et al. (2019) combines Convolutional Neural Networks (CNNs) with Fuzzy Time Series for short-term forecasting, while Khandelwal et al. (2015) merges ARIMA with ANNs for time series predictions. Similarly, Smyl (2020) integrates exponential smoothing with RNNs, and Weng et al. (2020) mixes Long Short-Term Memory (LSTM) with Light Gradient Boosting Machine (LightGBM), where LightGBM provides the interpretability.

During the research, it was not found studies on spirit drinks, the main topics were, in opposition: retail Alon et al., 2001; Fildes et al., 2022; Ma and Fildes, 2021; Punia et al., 2020,

fashion (Liu et al., 2013 and KALAOGLU et al., 2015) and spare parts (Ghobbar and Friend, 2003 and Eaves and Kingsman, 2004).

It is also to note, that as the company in study is a producer, its clients are other companies, that will resell the product. This implies intermediates between the company and the final consumer, as well as the data available being only sell-in (amount sold to these intermediate companies). With this observation, when compared to the literature, the studies found, are done with sales done to the final consumer.

2.3. External Factors

The need to identify external factors - variables that influence others - is not only a business need, as it is a way to enhance forecasting accuracy. As Abolghasemi et al., 2020 states: "Several endogenous and exogenous variables can influence the dynamics of demand, and hence a single statistical model that only consists of historical sales data is often insufficient to produce accurate forecasts".

Although the case study focuses on alcoholic drinks, this represents a particular area that is poorly covered about other factors. Therefore, to identify potential candidate factors, a broader research approach was employed, extending to the spirits industry as a whole. This approach acknowledges that factors influencing the broader spirits industry may produce varying results in this specific case, and their relevance will be tested accordingly.

2.3.1. Weather

Weather factors are proved on the literature to condition human behavior, as well as condition dislocations, affecting a broad amount of sectors, and even impacting the economy of the countries (Lazo et al., 2011).

Something as simple as the mood is influenced by the weather. Keller et al. (2005) observes an improvement on mood over pleasant weather during the spring (which the author describes as higher temperature or barometric pressure), and a decrease in mood during the summer on hotter weather. However, it is not just temperature that affects mood - sunlight also plays a role, showing a positive correlation (Murray et al., 2010). High humidity, on the other hand, suggests a state of enervation (Sanders and Brizzolara, 1982). Interestingly, mood also influences consumer behavior (Gardner and Hill, 1988), with negative moods leading to more rational choices, while positive moods result in more experiential decision-making.

The way a consumer behaves or perceives something, is also affected by the weather, warm temperatures increases the subject valuation of a product (Zwebner et al., 2014). During rainy weather, people will buy more expensive price items in a single order, and when temperature rises, it is expected they will buy fewer cheaper items (Tian et al., 2021), and yet on the subject of the basket size, it decreases on weather events (rain, snow, thunder and fog) (Moon et al., 2018), and cooperating with that, the consumer spending tends to increase with the increase of sunlight exposure (Murray et al., 2010), this same study finds a causal effect of sunlight on willingness to pay. On a study about restaurants complaints (Bujisic et al., 2019), it was demonstrated how it is more likely for customers to leave more negative comments when facing more

unpleasant weather that is observable, from where it was identified temperature, rain and barometric pressure, as the ones affecting the consumer behavior. Moving to mobile promotions, during a research on its effectiveness based on weather, it was found that on sunny weather the responses were higher and faster comparing to cloudy weather, and that the responses were lower and slower, on rainy weather. For non-alcoholic beverages, the purchase is done on an impulsive form, instead of planned (Štulec et al., 2019).

In terms of dislocations, precipitation and wind speed have effect on pedestrian flows, with a greater impact on winter months (Miranda-Moreno and Lahti, 2013). Adverse conditions like rain, extreme cold or hot temperatures, icy roads, heavy rain, dense fog or snow may discourage possible clients to dislocate, or even make it impossible to do so, if roads get blocked (Agnew and Thornes, 2007; Parnaudeau and Bertrand, 2018 and Moon et al., 2018). A similar pattern is observed in returning processes, with rain and extreme temperatures negatively affecting them (Hu et al., 2024). On rainy days, the number of visitors increases on shopping malls and decreases on street stores (Martínez-de-Albéniz and Belkaid, 2021).

Now exploring the impact of weather on sales, the earliest article encountered is Steele (1951), which analyzed the sales on a department store by considering the variables of precipitation amount, snow cover depth, temperature and wind velocity, concluding that 42% of the variance in store sales could be explained by weather factors. Murray et al. (2010) studied a retail store, and found the following variables could affect retail sales: temperature, humidity, snow fall, and especially sunlight. Arunraj and Ahrens (2016), also for retail, found that for food sales, snowfall and rainfall have a significant effect, while for fashion, snowfall has a significant effect, and temperature deviations a high effect. In this study, weather improved the fitness of the model by 4 to 6%. In the context of e-commerce, it was found rain, temperature and sunshine have an impact on daily sales, with a greater effect during summer, on weekends and on days with extreme weather (Steinker et al., 2017). The use of weather data in the previous study, led to the conclusion that it can reduce forecast error on average from 8.6% to 12.2%, with reductions of up to 50.6% on summer weekends. On a convenience store in China, it was found air quality having a negative impact on sales, and temperature having a positive one (Tian et al., 2021).

In Keleş et al. (2018), a study was conducted in the US across 52 markets, covering six liquid refreshment beverages (LRBs). The study established both long- and short-term relationships, revealing that rising temperatures lead to a yearly increase in demand for LRBs by approximately 0.21% (equivalent to 63 million gallons). This trend is further influenced by differences in demand during heat waves and cold waves, with the former leading to a significantly higher increase in weekly beverage demand.

Murray et al. (2010) analyzes non-alcoholic beverages and shows that weather sensitivity is not constant; it depends on the month, product category and store type. The study also indicates that for non-alcoholic beverages, temperature is the weather variable with strongest impact on sales. Additionally, it finds that during the summer months, from May to September, the correlation between sales and temperature is three times stronger than in the winter months.

In June and August, sales accelerate beyond a threshold temperature, while on the other summer months, the increase in sales is almost proportional to average daily temperature.

Regarding accuracy improvements, aside from the previously mentioned studies, Badorf and Hoberg (2020) reports forecast gains up to 7 days in advance. Similarly, Chan and Wahab (2024) demonstrates significant improvements in sales forecasting, accounting for up to an additional 47% of the variance in individual product sales and up to 56% in product categories, beyond what the baseline model explained. The study also identified the importance and influence on each specific weather feature.

2.3.2. Holidays

Holidays are days to celebrate and hang out with friends and family, which makes a perfect target to consume beverages, and as such to buy them. Hirche et al. (2021) denotes that public holidays influence short term sales, on a set of alcoholic beverages.

As a practical example, Zhang et al. (2020) found that cigarette sales on a Chinese company, have statistically significant positive pre-holiday effects. On Groene and Zakharov (2024) it is used three types of holidays: public holidays, local school holidays and regional school holidays, to help forecast the demand on food and beverages businesses. While applying exploratory data analysis to a Walmart dataset of sales, Shrieenidhi et al. (2024) finds a higher mean of sales for holidays.

2.3.3. Economic Indicators

Economic factors influence various sectors, including consumer behavior, which fluctuates depending on the economy. Consumers' purchasing power increases or decreases in line with economic conditions.

In the cosmetic industry, currency exchange fluctuations and seasonal factors have been found to significantly impact moisturizer sales, while inflation rates affect perfume sales (Khajehzadeh et al., 2022). A study on Chinese automobile sales found that the Consumer Confidence Index (CCI) and Consumer Price Index (CPI) have a causal relationship with automobile sales (Gao et al., 2018). Tourism demand is influenced by Gross Domestic Product (GDP) and exchange rates (Lin and Lee, 2013). Additionally, Kapoor and Ravi (2009) shows that an increase in interest rates leads to an immediate decline in consumer expenditure, while Ramkumar and Srinivasan, 2020 highlights how taxes on goods and services influence consumer behavior.

Looking at the retail sales, Bauerová et al. (2022) found influence of GDP, employment and gross wage.

In C.-H. Wang (2022), a study on sales forecasting applied to Taiwan's retail sector, three segments are examined: hypermarkets, supermarkets and convenience stores and claims they are "closely related to the economy condition of a country because they satisfy basic requirements in Maslow's hierarchy of needs". Overall, the Consumer Price Index, Retail Employment Population and real wage impact sales. Interestingly, unemployment rate, oil price and Consumer Confidence Index have no effect on sales, which contrasts with the earlier findings in C.-H. Wang (2022), where employment was identified as a factor influencing retail sales. For

specific sectors, the Wholesale Price Index is critical for convenience stores, while the Industry Price Index(IPI) and transportation freight are key for supermarkets. Seasonal factor, however, is only critical for hypermarkets.

Several months later, the lead author from the previous article, published a new study, similar, this time focusing on the United States of America, and analyzing three retailers (Walmart, Costco and Kroger), (C.-H. Wang and Gu, 2022). Overall, the Consumer Price Index and regular wage significantly influence on sales. Walmart is particularly affected by Product Price Index (PPI) and oil price, while it shares Gross Domestic Product as a critical factor with Costco. The Dow Jones Transportation Index is important for both Costco and Kroger, and Personal Consumption Expenditure influences both to Walmart and Kroger.

2.4. Consolidation

With all the intended subjects covered on the literature, now it is done a wrap up of the main aspects to takeaway and how it can relate to this project. Unfortunately, no relevant information was taken from the IWSR reports, regarding the products under study.

For the forecasting methods, it was concluded how machine learning methods, especially the deep learning ones, are expected to have a better performance than statistical ones, because of the capture of non linear patterns, but having in mind that its usage removes interpretability, so the usage of hybrid methods, as mentioned in the end of this matter, could be an interesting approach.

In terms of external factors, it was shown how weather affects a lot of circumstances of our lives, such as the mood, the consumer behavior and consequently sales. From the weather, it was observed the follow variables affecting sales: temperature, sunlight, humidity, precipitation, snow, wind velocity and air quality.

Holidays were also mentioned as having influence on sales, but as it is being used monthly data on this project, the granularity is not small enough to test it, leading it to not being used.

Lastly, from economic indicators, in terms of retail, which is the closest to current study, was found the follow variables influencing the sales: Gross Domestic Product, unemployment, wage, Consumer Price Index, Retail Employment Population, Wholesale Price Index, Industry Price Index, oil price, Personal Consumption Expenditure and Dow Jones Transport Index, which would have to be substituted to a similar index on Portugal, to make sense to use. On other sectors, was pointed: currency exchange, inflation rate, interest rate, Consumer Confidence Index and taxes.

CHAPTER 3

Methodology

This chapter outlines the steps and techniques applied, along with theoretical explanations. Broadly, the project follows three main phases: gaining business knowledge (Business Knowledge), preparing and exploring the data for further modeling (Exploratory Data Analysis), and modeling the data for predictions (Modeling). These phases, along with their sub-steps and key tasks, are illustrated in Figure 1.

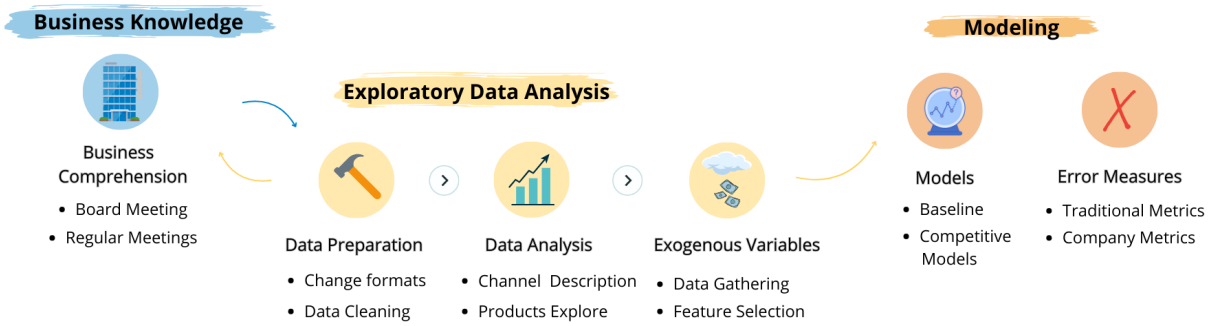


FIGURE 1. General Methodology Applied

3.1. Business Knowledge

Working with data requires to have knowledge about it. First, to be aware of the project, a presentation was read, and after that, some regular meetings would happen to have a better understanding of the needs of the company, and later to work with the data on Exploratory Data Analysis, for example, to understand the existence of null values in certain contexts.

Besides regular meetings, it happens a board meeting where were present the representative of the company, which would to the regular meetings, and who is responsible for planning and logistics, as well as a representative for each of the following departments: on-trade, off-trade, marketing, and finance.

In this meeting, shortly, it was possible to get deeper knowledge about:

- Their competitors, and how private labels are one of their biggest problems
- The national culture of consuming drinks outside, the reason why on-trade have way higher sales than off-trade
- Marketing initiatives, although there was no data given to work with
- Factors that, in their opinions could influence sales, and for such, the following were collected:
  - A lot of fragmentation - the fact that the product can be resold multiple times
  - Competitors

- Marketing actions and campaigns
- Intermediary impacts - this is related to the previously mentioned fragmentation. The example given was, that if a reseller discounted a competitor drink, it created a disadvantage for the company
- Tourism - from other countries comes other cultures, and by tourists ordering drinks that use not so common spiritual drinks in the country of visit, the drinks start to gain popularity, an example given was rum with coke
- Financial conditions of consumers - less money to spend, means a retraction on sales

3.2. Exploratory Data Analysis

As the image in chapter 3 suggests, this section covers three main topics: Data Preparation; Data Analysis and Exogenous Variables.

3.2.1. Data Preparation

Data Preparation has the goal of preparing the data to be analyzed, and is divided into two main tasks: Changing Formats - which implies the transformation of an excel file into a python *Data Frame* to work with *pandas* toolkit; and Data Cleaning, where null values are treated based on business logic.

3.2.2. Data Analysis

Data Analysis can be done after the data is prepared, and it has two main goals: describe off-trade channel and create a contextualization of the products' environment; and analyze individual product sales to search essentially for trends, seasonality, impact of COVID and behavior through time. Additionally, it is a studied process to make the products stationary and autocorrelations to capture existent correlations between a point in time, and past values.

3.2.3. Exogenous Variables

Exogenous Variables are candidate factors to help make predictions. On the Data Gathering phase, it was collected exogenous variables from the literature, and tried to find equal or similar ones for the Portugal context. Besides the variables from the literature, it was experimented house prices, given the housing crisis existent in Portugal, and as DJT (Dow Jones Transports index) was found in the literature, and Portugal has no transport index, it was used the main one - Portugal Stock Index (PSI).

Having a list of exogenous variables, doesn't mean that all support the predictions. The relevance of these variables in relation to each target SKU, is evaluated in this study through Granger causality, but before applying it, it is done a step of data processing, followed by a selection of features to evaluate causality, based on correlation.

Data processing involved transformations such as resampling, converting dates with month names into numeric *DateTime* variables, and removing extra rows on Excel files. After the cleaning process, correlations between the exogenous variables were calculated to address multicollinearity, by removing highly correlated features. This step was performed prior to testing

for causality to reduce the number of variables requiring stationarity transformations, and also to analyze causality. However, it could have been applied afterward as well.

Two variables are considered multicollinear when they exhibit a high correlation, meaning the information of one can be inferred from the other. Removing one variable from a highly correlated pair eliminates redundancy and reduces complexity in future models, as it leaves fewer parameters to estimate.

Correlation coefficients range from -1 to 1, where a value close to 0 indicates little to no correlation, meaning one variable does not affect the other. Values closer to 1 or -1 represent a strong correlation in either a positive or negative direction. A high positive correlation means that as one variable increases, the other also increases, while a high negative correlation means that an increase in one variable leads to a decrease in the other. To identify strong correlations, a common threshold of an absolute value of 0.7 was used. To determine which variables to remove, the following criteria were used: the number of variables with which each variable is highly correlated, domain knowledge, and variance.

After removing these variables, Granger causality was checked for remainders. Calculating the causality of exogenous variables on the target SKUs provides a measure of the influence these variables may have on predicting the target SKU. To evaluate this, Granger causality was applied, which is a hypothesis test assessing the null hypothesis that an exogenous variable does not Granger cause the target variable. If the test result is below a certain threshold, the null hypothesis is rejected, indicating that the exogenous variable has potential predictive power over the target. For this analysis, a more permissive significance threshold ($\alpha = 0.15$) was chosen.

3.3. Modeling

In this study, the SARIMA model was used as a baseline, and for competing models, were used SARIMAX and Extreme Gradient Boosting Machines (XGBMs or XGBoost). The baseline is the reference to compare with, when building other models, and understand if the predictions are good. In what follows, the employed models are briefly presented.

SARIMA is an extension of the ARIMA model that incorporates seasonal components to capture repeating patterns over time. ARIMA models consist of three parts: AutoRegressive (AR), which models the current value as a function of its past values; Integrated (I), which applies differencing to make the series stationary; and Moving Average (MA), which models the current value as a function of past forecast errors. SARIMA includes additional terms to model seasonality, as this was observed in the time series, on subsection 4.4.2. ARIMA and SARIMA models are particularly effective when dealing with relatively small datasets, provided the data exhibits patterns such as trends or seasonality.

SARIMAX is similar to SARIMA model, used in the baseline, but with the addition of exogenous variables, as indicated by the ”X” in SARIMAX. The goal is to improve performance by including these explanatory variables.

XGBMs is a machine learning algorithm, that builds multiple decision trees, in order to make predictions. Because of this, it is possible to infer features' importance in the model, which is helpful in the objective of finding factors affecting sales.

In terms of metrics, it was considered the Root Mean Squared Error (RMSE), which uses squared values to penalize bigger errors, and the root in the end to keep the value in the same scale as the original values. The following equation gives the formulae for this performance metric:

$$\text{RMSE}(y,\hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$$

From the company, two metrics are used to calculate the error, TISP and BIAS, which equates as follows:

$$TISP = \frac{\sum_{t=0}^{n} \hat{y}_t - \sum_{i=0}^{n} |\hat{y}_t - y_t|}{\sum_{i=0}^{n} \hat{y}_t}$$

$$BIAS = \frac{\sum_{t=0}^{n} (y_t - \hat{y}_t)}{\sum_{t=0}^{n} \hat{y}_t}$$

TISP measures the ability to accurately plan the product mix for a brand, while BIAS measures if the prevision was done below or above. The values are kept in the report, to serve a company need, but are not interpreted.

CHAPTER 4

Exploratory Data Analysis

With a fundamental understanding of the problem under study and after a review of the literature, it's time to get to know the data. This chapter describes the data provided by the company, the transformations applied to it, and the subsequent explorations conducted.

4.1. Data Description

The sales data covers a single brand and its GCBS (Group Codes Bottle Sizes), meaning that each SKU is a bottle size variation of a group code. The file is in format xlsx, a typical Excel sheet, containing monthly sales, covering the period from April 2013 to March 2024, in units of 9L boxes, through three tabs that differ in the context, being it:

- GCBS - Total monthly volume sales of a GCBS
- Channel - Monthly volume sales done through the channels. There were the channels: On-Trade, Off-Trade and Other
- Customer - Monthly volume sales done from each individual customer, anonymized, varying from channel.

For this analysis, it was discarded the channel Other, as it was not considered relevant for the business, remaining so, the principal channels: off-trade and on-trade. Given the project scope, only the tab Channel was considered for forecasting volume sales of products for both channels.

4.2. Data Transform

As described before, the data source is a sheet file, and the way the data is organized is not a proper way to work directly with it, so some adjustments had to be made. To perform the necessary processing, Python was used, specifically the *pandas* package.

From the header, was removed unnecessary information, as well as a column that would only have the values "Total Volumes". The date-time variable was presented by columns, instead of by row, so to simplify future processes, the table was transposed.

At first glance at the data in its original form, it was visible some products with blank values. After further analysis, it was concluded the existence of products that were released later and SKUs that were discontinued from sale. To deal with the null values, it was done the following:

- Find the release date, based on the first not null value found, and remove the null entries before that (Figure 2)

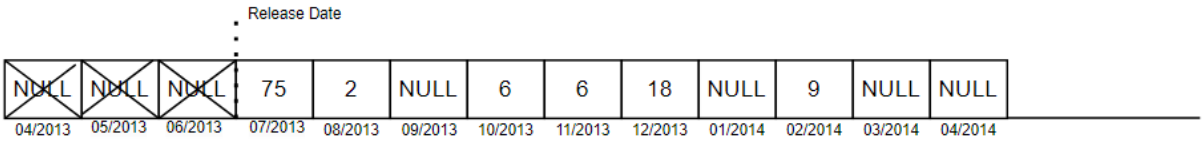


FIGURE 2. Removal of null values before release date

- Find the discontinued date, based on the first not null value found, in descending order of dates, and remove the latest null entries after that (Figure 3)

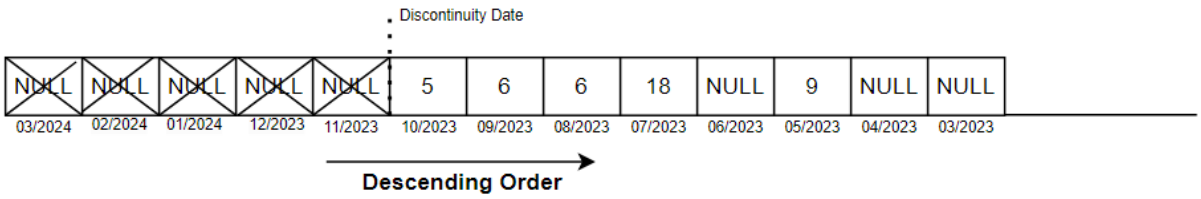


FIGURE 3. Removal of discontinuity period

- For intermediate null values, it was replaced with 0, as it was confirmed with the company as being absence of sales (Figure 4)

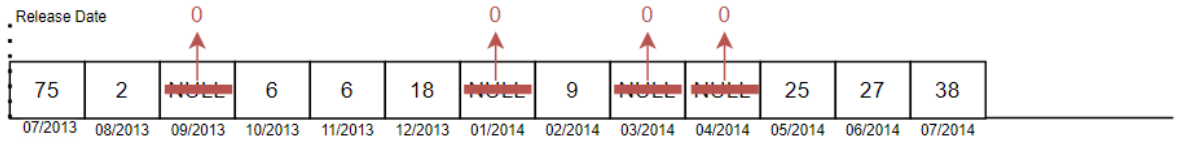


FIGURE 4. Replacement of intermediate null values with 0

Additionally, an anonymization of the products' names was done by mapping the group codes to an upper letter, varying from A to T.

4.3. Channel Analysis

Following a brief introduction to the data, an exploratory analysis was conducted to gain a deeper understanding, yielding descriptive statistics and visualizations that support comparisons and identify relationships. This analysis are divided by channel and products, for a smoother flow, starting with channel.

Power BI and Python scripts were used to perform this descriptive analysis.

4.3.1. General Insights

On the dataset is present a total of 20 group codes, resulting in 26 SKUs on the channel.

After addressing the blank values, as mentioned in section 4.2, numerous SKUs with many zeros were observed during this phase. This resulted in the time series having significantly fewer values to work with, which would impact the forecasting phase if we were to include all products. In Figure 5, it's possible to visualize the total number of observations on each SKU, as well as the correspondent weight of zeros, through percentage and remaining values to use.

Group Code	Bottle Size	Observations	Impact	Non Zero Values
E	100 CL	132		132
G	100 CL	132	-3,79%	127
Q	100 CL	132		132
E	6 CL	132	-54,55%	60
Q	6 CL	132		132
A	75 CL	132		132
E	75 CL	132		132
F	75 CL	132	-21,21%	104
Q	75 CL	132		132
S	75 CL	132	-37,12%	83
O	75 CL	117	-30,77%	81
T	75 CL	89	-32,58%	60
Q	50 CL	86	-60,47%	34
L	75 CL	73	-8,22%	67
N	75 CL	73	-9,59%	66
H	75 CL	66	-7,58%	61
B	100 CL	64	-82,81%	11
O	100 CL	62	-96,77%	2
K	75 CL	62	-77,42%	14
R	75 CL	62	-80,65%	12
I	75 CL	48		48
J	75 CL	48		48
C	25 CL	37		37
M	70 CL	32	-81,25%	6
D	75 CL	23		23
P	75 CL	12	-8,33%	11

FIGURE 5. Number of observations on each SKU and amount of non zero values.

4.3.2. Categories

Using a company portfolio image, products were categorized based on their type, but the process of linking them to the data had to be done manually. After completing the association, it was discovered that most products belonged to the Appetizer category. The number of products in each category is shown in Table 2.

Category	Number of Group Codes
Appetizer	8
Ready To Serve	4
Sparkling Wine	4
Riserva Speciale	3
Ready To Drink	1

TABLE 2. Distribution of group codes through categories

Appetizer not only represents the category with the most products in this study but also lead in sales volume, as shown in Figure 6, which illustrates the sales volume for each category.

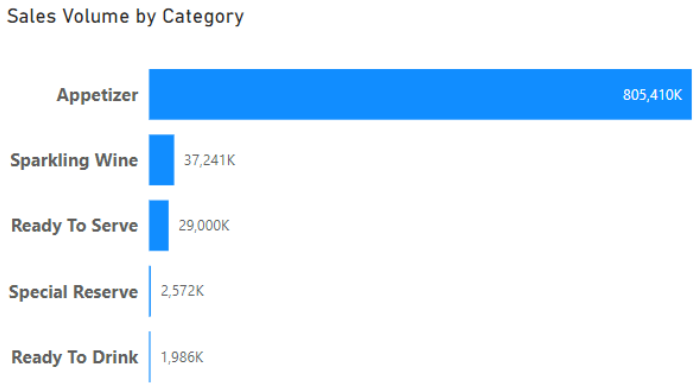


FIGURE 6. Volume of Sales on each category. Here is observable the high dominance on Appetizers.

The dominance of appetizers is not due to having the most products, but rather the exceptionally high sales volumes of group codes Q and E, which significantly surpass the others. This is illustrated in Figure 7 which shows the top five best-selling group codes on the channel. The first three group codes include the legend with their respective sales values, and for all, category is indicated by color, accompanied by the legend.

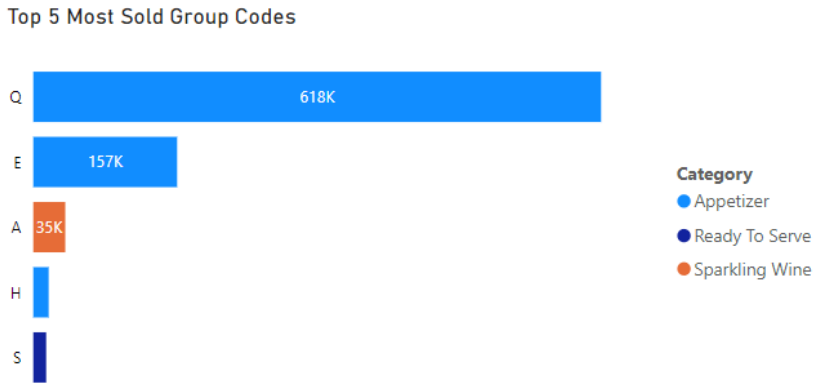


FIGURE 7. Top 5 most sold group codes, with total volume of sales and correspondent category, distinguished by color.

Among these five group codes, it's notable that three fall under the appetizer category. Sales for group code Q are significantly higher than the second best-selling group code, being nearly four times greater and accounting for almost 74% of the channel's total sales.

4.3.3. Bottle Size

Bottle sizes can vary in the following volumes: 6 CL; 50 CL; 75 CL and 100 CL. The distribution of sales for each bottle size can be found at Figure 8. From this graph, we can conclude that the majority of sales on this channel are made with 75CL bottles, accounting for 62.6%, followed by 100CL, 6CL, and 50CL bottles.

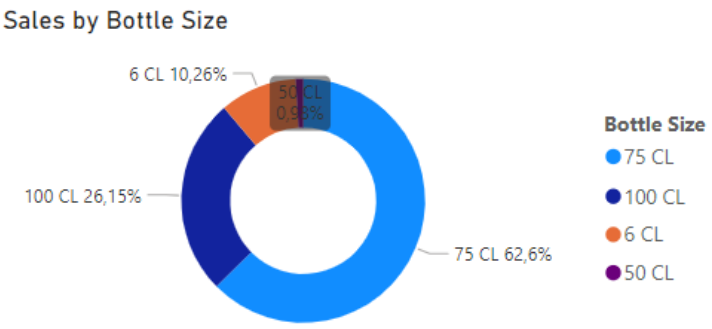


FIGURE 8. Relation of volume of sales between bottle sizes

4.3.4. Sales Behaviors

Over the years, the sales on off-trade had flutuacting between an average, as it can be seen in Figure 9.

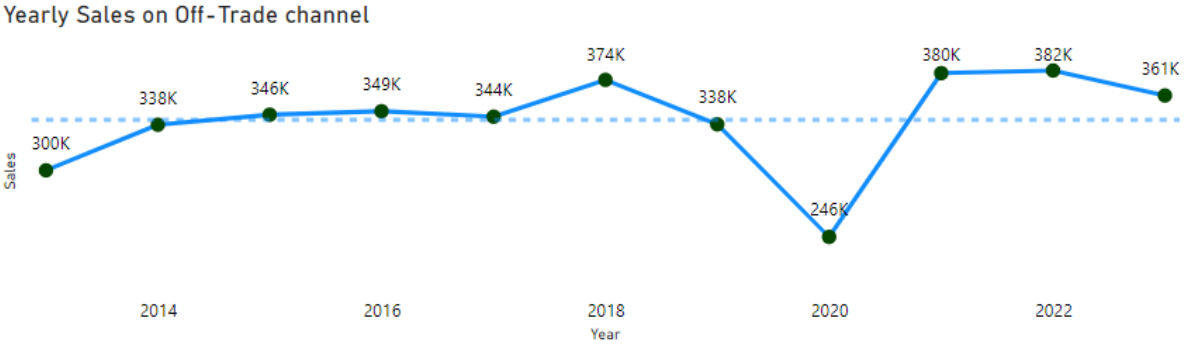


FIGURE 9. Yearly sales of channel

To analyze the monthly sales behavior, the same aggregation was employed, summing the volumes from all SKUs, as shown in Figure 10. The graph reveals a slow start in sales, with January recording the lowest sales of the year. Notable increases occur from June to August, when sales exceed the average monthly sales for the year. A similar trend is observed from October to December, although the relationship is less linear, with a sharp increase from November to December. While the analysis spans from the beginning of the year to the end, the final month significantly influences the first, as high expenses in December are driven by Christmas and New Year's Eve celebrations.

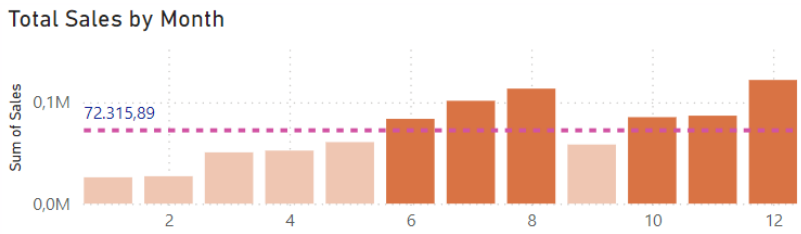


FIGURE 10. Total of monthly sales of channel, with an average line, where months with sales above it, are highlighted in a darker color.

4.4. Products Analysis

As established on chapter 1, the study for sales forecasting is only employed on top 5 best-selling SKUs by volume, and so, in conformity, individual product analysis are only applied to them. The products in question are: Q(75 CL), Q(100 CL), E (75 CL), Q(6 CL) and E(100 CL), whose sales amounts can be viewed and compared in Figure 11.

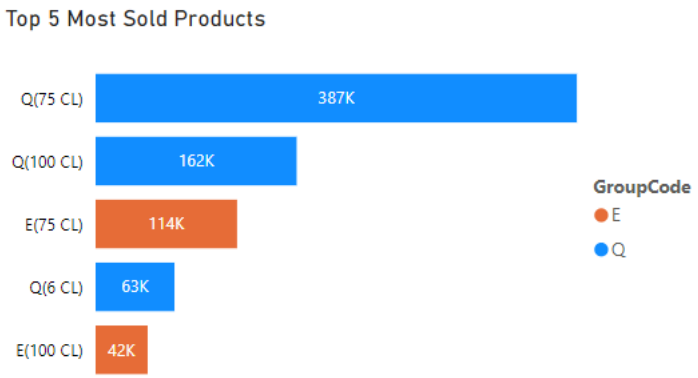


FIGURE 11. Top 5 best-selling products, which represent the targets

Examining the products, we can infer that they include only two group codes: Q and E, which correspond to the top-selling group codes, as shown in subsection 4.3.2, and both belong to the appetizer category.

To understand the products we have and compare them to identify possible relationships or similarities, all time series were first plotted in a single graph, as shown in Figure 12.

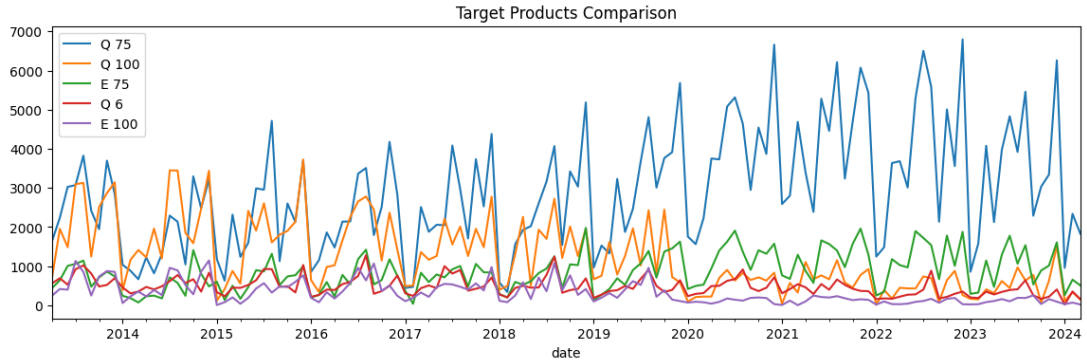


FIGURE 12. Combined Plot of Target Time Series

From this plot, it was observed that there is a significant gap between Q 75 and the other products, with all products exhibiting a similar seasonal pattern but differing in overall scale and trends.

Furthermore, it was calculated the correlations between them, as disposed in Figure 13.

	Q 75	Q 100	E 75	Q 6	E 100
Q 75	1.000000	0.036763	0.935810	0.349958	0.000970
Q 100	0.036763	1.000000	0.035718	0.644865	0.912795
E 75	0.935810	0.035718	1.000000	0.328136	0.000586
Q 6	0.349958	0.644865	0.328136	1.000000	0.566599
E 100	0.000970	0.912795	0.000586	0.566599	1.000000

FIGURE 13. Correlation Matrix between target products

From the correlation matrix, a strong correlation was observed between products Q 75 and E 75, as well as between Q 100 and E 100.

By comparing the value ranges shown in the boxplot displayed in Figure 14, a descending ladder-like pattern is evident from the first product to the last. The transition between the first and second products is notably more abrupt, as observed in previous graphics. However, starting from the second product onward, the differences appear to be more gradual. Descriptive statistics can be found in annex in Appendix A.

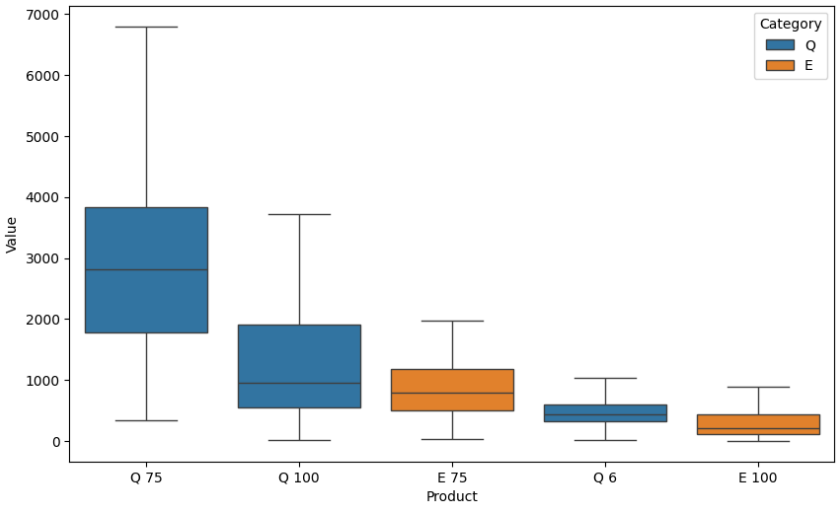


FIGURE 14. Boxplots Comparing the Distribution of Values Among Target Products.

4.4.1. Sales Evolution

To begin analyzing the products individually, a line plot of each product's sales evolution over time was created. The plot includes vertical bars marking each year to highlight potential seasonality, along with a shaded red area indicating the COVID period.

Starting with the top-selling product, Q 75, the plot in Figure 15, reveals an initial upward trend that gradually declines, with sales following an annual seasonal pattern. Notably, the impact of COVID on the sales of this product appears minimal.

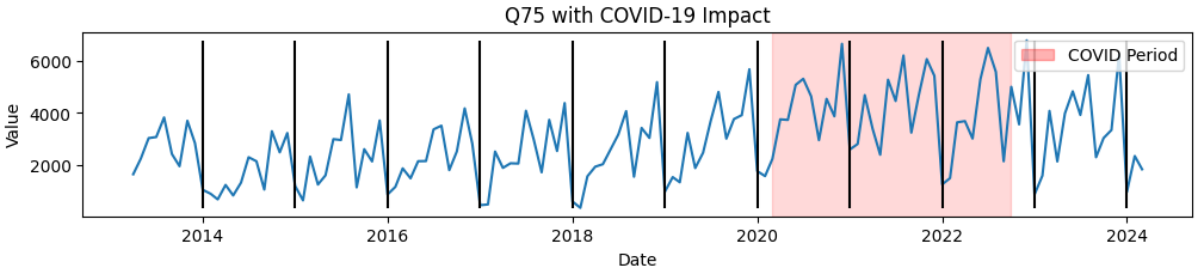


FIGURE 15. Evolution of Sales over Time for Product Q 75

The second best-selling product is Q 100. As shown in Figure 16, its sales exhibit a declining trend with some seasonal patterns. Unlike the 75 CL version, the 100 CL sales were significantly impacted by COVID, leading to a sharp decline.

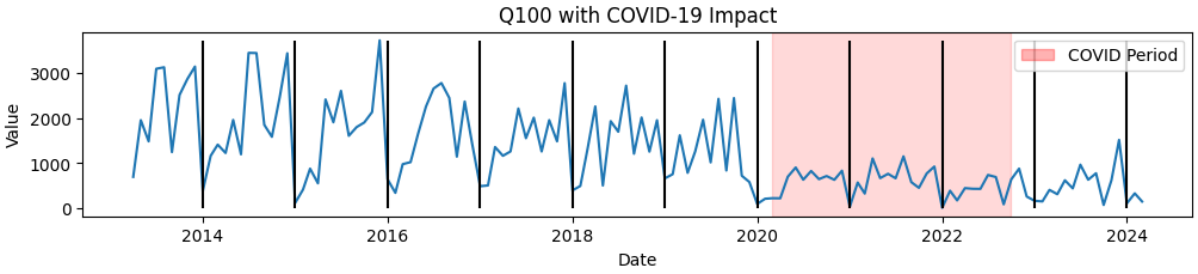


FIGURE 16. Evolution of Sales over Time for Product Q 100

The third-most sold product is E 75. As shown in Figure 17, both trends and seasonality are evident. Despite the COVID period, sales remained strong.

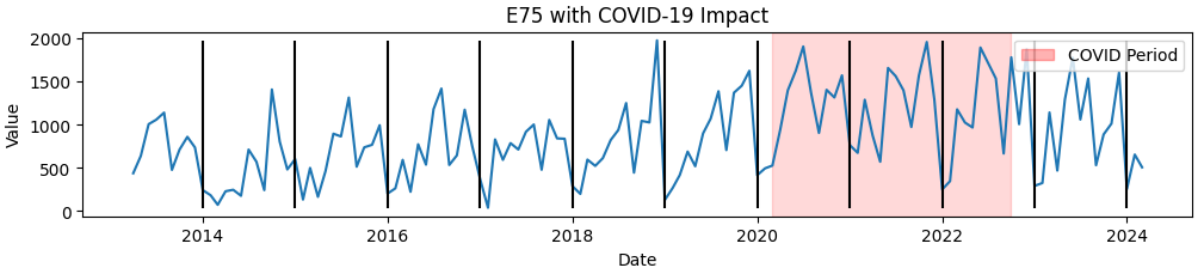


FIGURE 17. Evolution of Sales over Time for Product E 75

The fourth-most sold product is Q 6. As shown in Figure 18, the series exhibits a downward trend along with seasonality. During the COVID period, a noticeable decline in sales is evident.

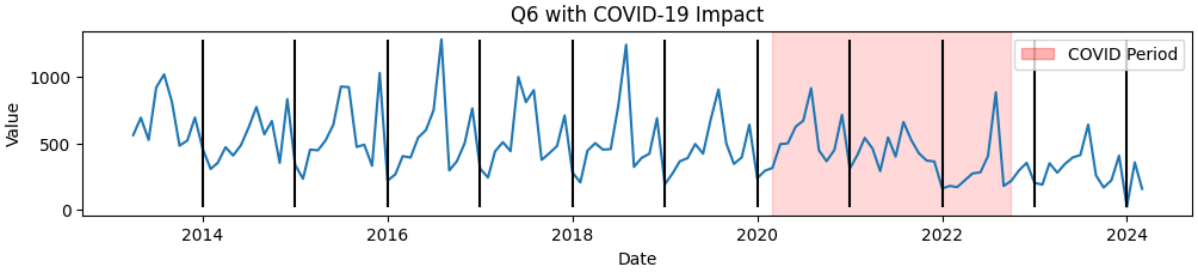


FIGURE 18. Evolution of Sales over Time for Product Q 6

The fifth-most sold product, and the last one in this study, is E 100. As shown in Figure 19, there appears to be no discernible trend, but the plot suggests the possibility of seasonality. During the COVID period, there was a significant drop in sales that has not yet recovered.

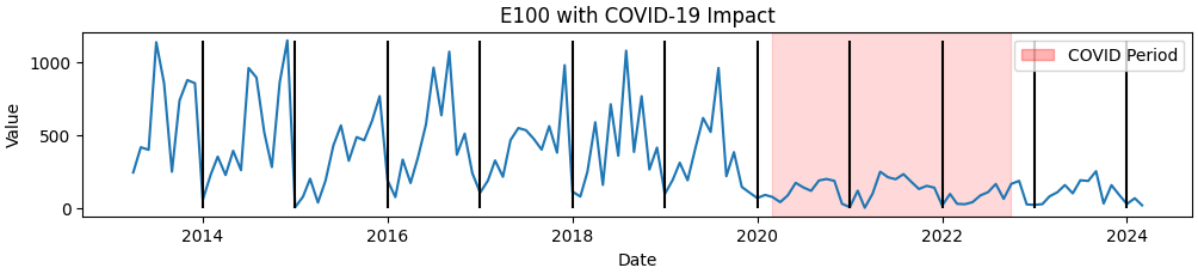


FIGURE 19. Evolution of Sales over Time for Product E 100

At this stage, it was also explored whether applying logarithmic transformations would bring the distribution of the data closer to a normal distribution, aiming to reduce asymmetry and variance. For the five products analyzed, the conclusion was consistent: the logarithmic transformation actually accentuated the asymmetry, as shown in Appendix B. Therefore, the logarithmic transformation was not retained.

4.4.2. Stationary Process

To investigate the presence of trend and seasonality in the series, a decomposition was performed, and the resulting components are presented in Appendix C. Additionally, the Augmented Dickey-Fuller (ADF) test was applied to assess stationarity. The ADF test is a hypothesis test where the null hypothesis posits the existence of a unit root, indicating that the series is non-stationary. Using a significance level (denoted by α) of 0.05, the p-value obtained from the test must be less than α to reject the null hypothesis, thereby concluding that the series is stationary. The conclusions regarding the presence of trend, seasonality, and stationarity are summarized in Table 3.

Product	**Trend**	**Seasonality**	**Raw ADF p-value**	**Differentiations**	**Final ADF p-value**
Q 75	Stochastic Trend	Yes	0.9628	1 lag for trend 12 lags for seasonality	7.0984e-6
Q 100			0.7191		0.0078
E 75			0.9646		0.0002
Q 6			0.8353		0.0053
E 100			0.5547		7.1275e-5

TABLE 3. Analysis of Trend, Seasonality, and Differencing in the Product Series

4.4.3. Autocorrelations

With the series made stationary, the correlations between a data point and its previous values were analyzed. This was done by computing the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF). The ACF measures the overall correlation between a point and a past value, capturing both direct and indirect effects of intermediate lags. In contrast, the PACF isolates the direct correlation between a point and a specific lag, removing the influence of the intermediate values, providing the partial correlation at that lag.

The purpose of computing these correlations is not only to understand how a point relates to its past values, but also to guide the modeling process for certain time series models, such as ARMA models and their variations, which are applied in this study.

Across all five products, consistent behaviors were observed in the analyses, leading to the same conclusions. The analyses, detailed in Appendix D, show that the ACF revealed a significant peak at lag 1 followed by a sharp decline. Meanwhile, the PACF displayed a more gradual decay, with three significant peaks identified for product Q 100 and two for the other products. This relationship between the ACF and PACF results supports the conclusion that there is an MA(1) component, indicating a Moving Average with a lag of 1, where the parameter 1 corresponds to the number of significant lags identified in the ACF. The number of peaks identified in the PACF corresponds to the possible lag values to consider for the AR component: from 1 to 3 for product Q 100, and from 1 to 2 for the other products.

4.5. Exogenous Variables

The identification of exogenous variables was guided by the findings from the literature review and the availability of relevant data for Portugal. Additionally, was tested house prices and Portugal Stock Index (PSI).

In the weather domain, the following variables were considered: maximum temperature felt, the minimum temperature felt, the average temperature felt, humidity, precipitation, probability of precipitation, wind gust, wind speed, cloud cover, and solar radiation. Given the available data, weather information specific to Lisbon was used for the analysis.

In the economic field, the variables tested were unemployment, CPI, CCI, GDP, PPI, IPI, PSI, house price, interest rate, oil price, and transportation distance (the total distance traveled by cargo trucks).

These variables were also processed, involving transformations like resampling to a monthly frequency to align with the sales frequency; converting dates with month names into numeric *DateTime* variables and removing extra rows on excel files. The methods used to resample the variables that require it are listed in Table 4.

Variable	**Original Frequency**	**Resampling Method**
GDP	Quarterly	Linear interpolation
PSI	Daily	Downloaded as monthly (source uses last value of the month)
House Price	Quarterly	Linear interpolation
Oil Price	Daily	First value of each month
Transportation Distance	Quarterly	Linear interpolation
Weather Variables	Daily	Average

TABLE 4. Resampling Methods Applied to Necessary Variables

4.5.1. Feature Reduction

Correlations were first calculated within each domain separately, and then combined to simplify the analysis. This resulted in individual correlation matrices for weather and economic variables, followed by a general matrix that combines both domains, already filtered for multicollinearity, along with the transport distance variable. To identify strong correlations, a common threshold of an absolute value of 0.7 was used

4.5.1.1. *Weather Variables* Beginning with the weather variables, their correlation matrix is displayed in Figure 20. The resultant strong correlations are further identified in Table 5, which also includes the count of strong correlations for the variable in the "Var 1" column.

	temperature	humidity	precip	precipprob	windgust	windspeed	cloudcover	solarradiation
temperature	1.000000	-0.662049	-0.464156	-0.658600	0.019929	0.303476	-0.581301	0.752063
humidity	-0.662049	1.000000	0.643730	0.780311	-0.077260	-0.249444	0.745748	-0.730789
precip	-0.464156	0.643730	1.000000	0.797091	0.257629	0.045810	0.600799	-0.558307
precipprob	-0.658600	0.780311	0.797091	1.000000	0.167634	-0.102134	0.772053	-0.693621
windgust	0.019929	-0.077260	0.257629	0.167634	1.000000	0.684293	0.037705	0.182539
windspeed	0.303476	-0.249444	0.045810	-0.102134	0.684293	1.000000	-0.106029	0.507779
cloudcover	-0.581301	0.745748	0.600799	0.772053	0.037705	-0.106029	1.000000	-0.546400
solarradiation	0.752063	-0.730789	-0.558307	-0.693621	0.182539	0.507779	-0.546400	1.000000

FIGURE 20. Correlations Between Weather Variables

Var 1	Var 2	Coefficient Value	Counts for Var 1
temperature	solarradiation	0.7521	1
humidity	precipprob	0.7803	3
humidity	cloudcover	0.7457	
humidity	solarradiation	-0.7308	
precip	precipprob	0.7971	1
precipprob	humidity	0.7803	3
precipprob	precip	0.7971	
precipprob	cloudcover	0.7721	
cloudcover	humidity	0.7457	2
cloudcover	precipprob	0.7721	
solarradiation	temperature	0.7520	2
solarradiation	humidity	-0.7308	

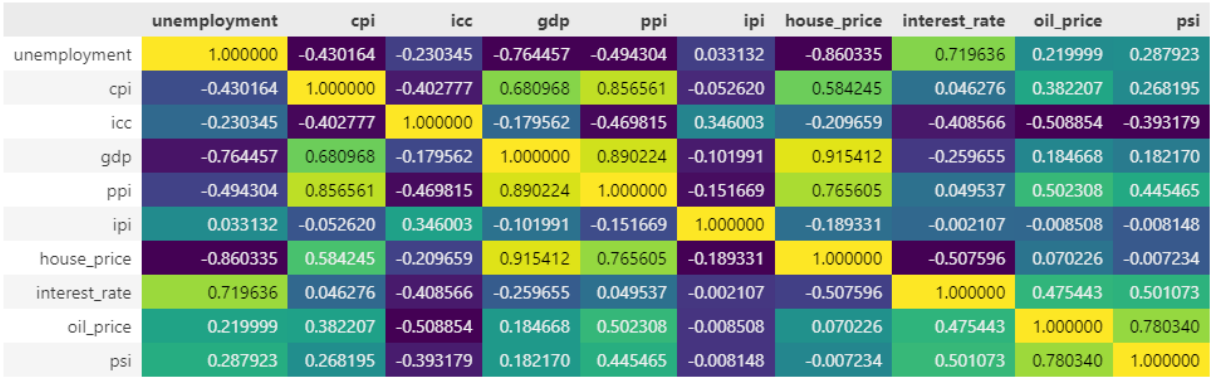
TABLE 5. Strong Correlations Present In Weather Variables

To determine which variables to remove, the following criteria were used: the number of variables with which each variable is highly correlated, domain knowledge, and variance.

Given this, humidity was removed due to its high correlation with three variables. Similarly, the probability of precipitation was excluded because it is redundant with precipitation and has

three strong correlations. Consequently, only the strong correlation between *solarradiation* and temperature remains, with a preference for temperature based on findings from the literature review. This leads to the following weather variables being selected: temperature, precipitation, wind gust, wind speed and cloud cover.

4.5.1.2. *Economic Variables* Moving to economic variables, it was obtained the correlation matrix present in Figure 21, where the same criteria was used, so to support it, there is also a table for strong correlations in Table 6.



	unemployment	cpi	icc	gdp	ppi	ipi	house_price	interest_rate	oil_price	psi
unemployment	1.000000	-0.430164	-0.230345	-0.764457	-0.494304	0.033132	-0.860335	0.719636	0.219999	0.287923
cpi	-0.430164	1.000000	-0.402777	0.680968	0.856561	-0.052620	0.584245	0.046276	0.382207	0.268195
icc	-0.230345	-0.402777	1.000000	-0.179562	-0.469815	0.346003	-0.209659	-0.408566	-0.508854	-0.393179
gdp	-0.764457	0.680968	-0.179562	1.000000	0.890224	-0.101991	0.915412	-0.259655	0.184668	0.182170
ppi	-0.494304	0.856561	-0.469815	0.890224	1.000000	-0.151669	0.765605	0.049537	0.502308	0.445465
ipi	0.033132	-0.052620	0.346003	-0.101991	-0.151669	1.000000	-0.189331	-0.002107	-0.008508	-0.008148
house_price	-0.860335	0.584245	-0.209659	0.915412	0.765605	-0.189331	1.000000	-0.507596	0.070226	-0.007234
interest_rate	0.719636	0.046276	-0.408566	-0.259655	0.049537	-0.002107	-0.507596	1.000000	0.475443	0.501073
oil_price	0.219999	0.382207	-0.508854	0.184668	0.502308	-0.008508	0.070226	0.475443	1.000000	0.780340
psi	0.287923	0.268195	-0.393179	0.182170	0.445465	-0.008148	-0.007234	0.501073	0.780340	1.000000

FIGURE 21. Correlations Between Economic Variables

Var 1	**Var 2**	**Coefficient Value**	**Counts for Var 1**
unemployment	gdp	-0.7645	3
unemployment	house_price	-0.8603	
unemployment	interest_rate	0.7196	
cpi	ppi	0.8566	1
gdp	unemployment	-0.7645	3
gdp	ppi	0.8902	
gdp	house_price	0.9154	
ppi	cpi	0.8566	3
ppi	gdp	0.8902	
ppi	house_price	0.7656	
house_price	unemployment	-0.8603	3
house_price	gdp	0.9154	
house_price	ppi	0.7656	
interest_rate	unemployment	0.7196	1
oil_price	psi	0.7803	1
psi	oil_price	0.7803	1

TABLE 6. Strong Correlations Present In Economic Variables

Following the criteria, the focus was on the variables with the highest number of strong correlations. Unemployment, GDP, PPI, and house price each had 3 strong correlations. The

decision was made to remove PPI first, as the literature suggests it only affects specific cases. This removal left unemployment with 3 strong correlations, so it was the next variable to be excluded. After these removals, only two strong relationships remained: between GDP and house price, where GDP was retained due to its relevance in the literature, and the relationship between oil price and PSI, where PSI was chosen to remain due to its higher variance. This leads to selecting the following economic variables: CPI, ICC, GDP, IPI, interest rate, and PSI.

4.5.1.3. *Combined Variables* Finally, both sets of variables were combined, along with transport distance, and the correlation between all variables was calculated. The results are shown in Figure 22. This time, no further strong correlations were identified, so the final list of variables to test for causality on the products consists of: temperature, precipitation, wind gust, wind speed, cloud cover, CPI, ICC, GDP, IPI, interest rate, PSI, and transport distance.

	temperature	precip	windgust	windspeed	cloudcover	cpi	icc	gdp	ipi	interest_rate	psi	transport_distance
temperature	1.000000	-0.464156	0.019929	0.303476	-0.581301	0.048128	0.006211	0.005770	0.011194	0.023213	-0.086801	-0.266483
precip	-0.464156	1.000000	0.257629	0.045810	0.600799	-0.041873	-0.023359	-0.060472	0.059183	0.054228	-0.074984	0.128949
windgust	0.019929	0.257629	1.000000	0.684293	0.037705	-0.399352	0.375915	-0.431036	0.200058	-0.065003	-0.236945	0.340681
windspeed	0.303476	0.045810	0.684293	1.000000	-0.106029	-0.071703	0.065330	-0.130599	0.127539	0.197322	0.115802	0.131103
cloudcover	-0.581301	0.600799	0.037705	-0.106029	1.000000	0.041165	-0.153250	0.013157	0.075460	0.040595	0.191596	0.186368
cpi	0.048128	-0.041873	-0.399352	-0.071703	0.041165	1.000000	-0.402777	0.680968	-0.052620	0.046276	0.268195	-0.315647
icc	0.006211	-0.023359	0.375915	0.065330	-0.153250	-0.402777	1.000000	-0.179562	0.346003	-0.408566	-0.393179	0.117353
gdp	0.005770	-0.060472	-0.431036	-0.130599	0.013157	0.680968	-0.179562	1.000000	-0.101991	-0.259655	0.182170	-0.692249
ipi	0.011194	0.059183	0.200058	0.127539	0.075460	-0.052620	0.346003	-0.101991	1.000000	-0.002107	-0.008148	0.261522
interest_rate	0.023213	0.054228	-0.065003	0.197322	0.040595	0.046276	-0.408566	-0.259655	-0.002107	1.000000	0.501073	0.403114
psi	-0.086801	-0.074984	-0.236945	0.115802	0.191596	0.268195	-0.393179	0.182170	-0.008148	0.501073	1.000000	0.065783
transport_distance	-0.266483	0.128949	0.340681	0.131103	0.186368	-0.315647	0.117353	-0.692249	0.261522	0.403114	0.065783	1.000000

FIGURE 22. Correlations Between Weather Variables

4.5.2. Causalities

Calculating the causality of exogenous variables on the target SKUs provides a measure of the influence these variables may have on predicting the target SKU. To evaluate this, Granger causality was applied, which is a hypothesis test where if the test result is below a certain threshold, the null hypothesis is rejected, indicating that the exogenous variable has potential predictive power over the target - more theorethical details on section 3.2. For this analysis, a more permissive significance threshold ($\alpha = 0.15$) was chosen. Before calculating the causality, the exogenous variables were differed each, until it got stationary. The Granger causality test results are presented in Figure 23, and the associated exogenous variables to each SKU, based on the threshold, are present on Table 7.

	temperature	precipitation	windgust	windspeed	cloudcover	unemployment	icc	ppi	ipi	psi	transport_distance
Q 75	0.600800	0.265500	0.090800	0.119100	0.045600	0.019100	0.042300	0.487600	0.138900	0.068700	0.342700
Q 100	0.010700	0.156600	0.145600	0.268700	0.280400	0.173500	0.355000	0.759000	0.364000	0.259300	0.000300
E 75	0.569700	0.368400	0.240700	0.304100	0.461800	0.007000	0.001500	0.510800	0.031400	0.015700	0.256700
Q 6	0.172600	0.118800	0.093800	0.029200	0.075600	0.055300	0.078800	0.111700	0.060200	0.010000	0.309700
E 100	0.029900	0.023200	0.153800	0.227100	0.110300	0.005300	0.225100	0.756900	0.210600	0.458100	0.038600

FIGURE 23. Causality Matrix Between Exogenous Variables and SKUs

	Q 75	Q 100	E 75	Q 6	E 100
temperature		x			x
precipitation				x	x
windgust	x	x		x	
windspeed	x			x	
cloudcover	x			x	x
unemployment	x		x	x	x
icc	x		x	x	
ppi				x	
ipi	x		x	x	
psi	x		x	x	
transport_distance		x			x

TABLE 7. Variables Associated with Each SKU Based on Causality Analysis

CHAPTER 5

Modeling

With the problem stated, and a better understanding of the data on hold, it is now possible to start modeling the data and find answers to the problem in hands.

This phase begins with the baseline models, developed using SARIMA, and then progresses to the experimental models, including SARIMAX and Gradient Boosting Machines.

5.1. Baseline (SARIMA)

Before testing different models, it is crucial to establish a baseline to provide reference values for comparison with the results of other models, allowing for the evaluation of their performance. In this study, the baseline model chosen is SARIMA.

To identify the optimal parameters for SARIMA, the *auto-arima* function from *pmdarima* was used, specifying potential values for the AR and MA components based on the analysis detailed in subsection 4.4.3. For each product, a separate model was built for each prediction horizon (3, 6, and 12 months). In addition to fitting the SARIMA model suggested by *auto-arima*, a recursive forecast was also generated based on the selected model. Model evaluation was performed using Root Mean Squared Error (RMSE) as the primary metric, with Mean Absolute Percentage Error (MAPE) included to provide a relative perspective on error. A 4-fold cross-validation (CV) was also applied to assess the model's generalization capability. The best model was selected based on the RMSE obtained from cross-validation, as it quantifies the generalization potential. The best model for each specified period is highlighted in the result tables to make it easier to identify.

Starting with the first product, Q 75, we can observe that for the 3 and 6-month horizons, the direct forecast presents the best results. This can be explained by the seasonality pattern in the time series and the short-term predictability power of the SARIMA model. For the 12-month prediction window, the recursive forecast achieved better results, as shown in Table 8, being a longer interval SARIMA loose predictability power and a rolling window 1-point recursive rule is beneficial.

Model	Forecast Period	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMA(1,1,1)(0,1,2)[12]	3	926.69	920.27	1708.00	0.525	0.009
SARIMA(1,1,1)(0,1,2)[12] Recursive	3	951.40	945.14		0.522	-0.007
SARIMA(1,1,1)(1,1,1)[12]	6	1020.48	948.96	2960.25	0.873	-0.750
SARIMA(1,1,1)(1,1,1)[12] Recursive	6	1053.67	976.26		0.688	-0.066
SARIMA(1,1,1)(1,1,1)[12]	12	1024.06	1117.83	3364.63	0.781	-0.135
SARIMA(1,1,1)(1,1,1)[12] Recursive	12	971.97	1002.26		0.758	-0.057

TABLE 8. Baseline Results for Product Q 75

For product Q 100, the recursive approach yielded better results in cross-validation for all cases; however, for the 3- and 6-month periods, the non-recursive RMSE showed lower error, as presented in Table 9.

Model	Forecast Period	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMA(0,1,1)(0,1,1)[12]	3	143.28	465.79	197.56	-0.079	0.516
SARIMA(0,1,1)(0,1,1)[12] Recursive	3	148.79	442.06		0.030	0.291
SARIMA(1,1,1)(0,1,1)[12]	6	472.16	443.78	468.56	0.799	-0.724
SARIMA(1,1,1)(0,1,1)[12] Recursive	6	507.89	418.06		0.006	0.291
SARIMA(1,1,1)(0,1,1)[12]	12	619.77	556.12	547.22	-5.751	6.337
SARIMA(1,1,1)(0,1,1)[12] Recursive	12	471.45	480.94		-0.096	0.629

TABLE 9. Baseline Results for Product Q 100

For product E 75, the recursive prediction was more effective for the 6 and 12-months horizons, as presented in Table 10.

Model	Forecast Period	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMA(0,1,1)(0,1,1)[12]	3	294.55	362.73	474.17	0.479	-0.021
SARIMA(0,1,1)(0,1,1)[12] Recursive	3	320.08	370.96		0.453	-0.051
SARIMA(0,1,1)(0,1,1)[12]	6	360.23	368.09	823.00	0.858	-0.770
SARIMA(0,1,1)(0,1,1)[12] Recursive	6	355.83	354.42		0.670	-0.119
SARIMA(0,1,1)(0,1,1)[12]	12	416.79	406.80	968.13	0.711	-0.210
SARIMA(0,1,1)(0,1,1)[12] Recursive	12	372.13	358.88		0.698	-0.107

TABLE 10. Baseline Results for Product E 75

For product Q 6, the recursive prediction was more effective for 6-month and 12-month period, as presented in Table 11.

Model	Forecast Period	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMA(2,1,0)(2,1,0)[12]	3	142.44	86.77	176.89	0.202	0.109
SARIMA(2,1,0)(2,1,0)[12] Recursive	3	164.20	94.80		0.099	0.070
SARIMA(2,1,0)(0,1,1)[12]	6	129.07	122.47	220.40	0.654	-0.440
SARIMA(2,1,0)(0,1,1)[12] Recursive	6	130.89	106.54		0.559	0.178
SARIMA(2,1,0)(2,1,0)[12]	12	102.74	181.83	303.62	0.752	-0.144
SARIMA(2,1,0)(2,1,0)[12] Recursive	12	103.27	126.71		0.740	-0.019

TABLE 11. Baseline Results for Product Q 6

For product E 100, recursive forecasting proved to be the best approach for all periods, as shown in Table 12.

Model	Forecast Period	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMA(0,1,1)(0,1,1)[12]	3	44.54	98.77	40.22	-7.597	8.424
SARIMA(0,1,1)(0,1,1)[12] Recursive	3	42.30	90.77		-2.059	2.159
SARIMA(0,1,1)(0,1,1)[12]	6	62.84	119.47	67.67	0.785	-0.732
SARIMA(0,1,1)(0,1,1)[12] Recursive	6	67.72	93.03		0.165	0.034
SARIMA(1,1,1)(0,1,1)[12]	12	168.02	176.45	117.61	5.355	-4.355
SARIMA(1,1,1)(0,1,1)[12] Recursive	12	98.88	144.50		-0.034	0.494

TABLE 12. Baseline Results for Product E 100

5.2. SARIMAX

The first model tested is SARIMAX, similar to the SARIMA model used in the baseline but with the addition of exogenous variables, as indicated by the ”X” in SARIMAX. The goal is to improve performance by including these explanatory variables.

Although a causality study was conducted, the variables may still lack a significant effect on the model. Therefore, a model was built for each product with one variable at a time, and a Z-test p-value was evaluated to assess the variable's impact on prediction accuracy. An α level of 0.10 was used, indicating that p-values below this threshold suggest the variable has a meaningful effect on the prediction.

For product Q 75, the variable 'icc' was identified as the only significant predictor for the 3-month and 6-month periods, with a p-value of 0.064 and 0.066, respectively, while 'gdp' was the sole impactful variable for the 12-month period, with a p-value of 0.056. Recursive forecasting showed improved results only for the 12-month period, as presented in Table 13. All p-values from the Z-test are presented in annex in Table 31.

Model	Forecast Period	Exogenous	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMAX(1, 1, 1)x(0, 1, [1, 2], 12)	3	icc	856.90	859.03	1708.00	0.504	0.111
SARIMAX(1, 1, 1)x(0, 1, [1, 2], 12) Recursive	3	icc	862.01	950.92		0.508	0.090
SARIMAX(1, 1, 1)x(1, 1, 1, 12)	6	icc	968.78	1012.93	2960.25	0.728	-0.098
SARIMAX(1, 1, 1)x(1, 1, 1, 12) Recursive	6	icc	1412.60	1231.94		0.485	0.142
SARIMAX(1, 1, 1)x(1, 1, 1, 12)	12	gdp	1142.43	1232.98	3364.63	0.770	-0.175
SARIMAX(1, 1, 1)x(1, 1, 1, 12) Recursive	12	gdp	941.49	1207.09		0.755	-0.001

TABLE 13. SARIMAX Results for Product Q 75

For product Q 100, 'gdp' was identified as the only significant predictor for the 3-month and 6-month periods, with p-values of 0.002 and 0.001, respectively, while 'temperature' was the sole impactful variable for the 12-month period, with a p-value of 0.023. Recursive forecasting showed improved results only for the 6-month period, as shown in Table 14. All p-values from the Z-test are presented in annex in Table 32.

Model	Forecast Period	Exogenous	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMAX(3, 1, 1)x(0, 1, 1, 12)	3	gdp	361.25	455.54	197.56	3.337	-2.337
SARIMAX(3, 1, 1)x(0, 1, 1, 12) Recursive	3	gdp	308.75	388.62		-0.632	0.283
SARIMAX(3, 1, 1)x(0, 1, 1, 12)	6	gdp	603.13	426.71	468.56	-2.470	2.903
SARIMAX(3, 1, 1)x(0, 1, 1, 12) Recursive	6	gdp	632.88	477.65		0.178	-0.154
SARIMAX(3, 1, 0)x(0, 1, [1], 12)	12	temperature	477.91	491.25	547.22	0.369	-0.090
SARIMAX(3, 1, 0)x(0, 1, [1], 12) Recursive	12	temperature	1512.61	1424.16		0.297	-0.703

TABLE 14. SARIMAX Results for Product Q 100

Product E 75 stood out as the only product with multiple influential variables. For the 3-month period, both GDP and PSI were significant, with p-values of 0.002 and 0.017, respectively. In the 6- and 12-month periods, GDP lost significance and was replaced by ICC, with p-values of 0.030 and 0.001, while PSI remained significant, with p-values of 0.013 and 0.008, respectively. Since two significant variables were identified per period, three models were developed for each period: one with both variables and one with each variable separately, totaling 18 models. Recursive forecasting was most effective for the 3- and 6-month periods with both variables, whereas in the 12-month period, a single-variable model using 'icc' performed best without recursive forecasting. Results for the best models are shown in Table 15, and all p-values from the Z-test are listed in annex in Table 33.

Model	Forecast Period	Exogenous	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	3	gdp psi	323.02	406.67	474.17	0.520	-0.204
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	3	gdp psi	319.80	345.29		0.490	-0.167
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	6	psi icc	340.89	339.55	823.00	0.700	-0.187
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	6	psi icc	362.06	321.31		0.620	-0.045
SARIMAX(1, 1, 2)x(0, 1, [1], 12)	12	icc	328.42	331.28	968.12	0.679	0.058
SARIMAX(1, 1, 2)x(0, 1, [1], 12) Recursive	12	icc	514.32	466.82		0.275	0.576

TABLE 15. SARIMAX Results for Product E 75

For product Q 6, the variable 'psi' was found to influence predictions in the 3-month and 12-month periods, with p-values of 0.071 and 0.084, respectively. For the 6-month period, however, no variable reached statistical significance, so 'gdp,' which had the lowest AIC, was tested instead. The best model for the 6-month period is highlighted in a distinct tone due to this unique selection. Recursive forecasting showed improved results in the 3-month period, as detailed in Table 16. All p-values from the Z-test are summarized in annex in Table 34.

Model	Forecast Period	Exogenous	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	3	psi	171.34	78.55	176.89	-0.204	0.856
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	3	psi	151.11	73.78		0.158	0.166
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	6	gdp*	130.46	122.61	220.40	0.591	0.350
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	6	gdp*	119.47	145.33		0.622	0.143
SARIMAX(2, 1, 1)x(2, 1, [], 12)	12	psi	89.91	127.91	303.62	0.760	0.124
SARIMAX(2, 1, 1)x(2, 1, [], 12) Recursive	12	psi	298.06	207.70		0.529	-0.423

TABLE 16. SARIMAX Results for Product Q 6. For the 6-month period, the variable with the lowest AIC was tested, as no significant variables were identified.

For product E 100, no significant variables were identified for predictions across all time periods. As a result, the variable with the lowest AIC, 'cloudcover,' was tested for all three periods. However, the recursive forecast performed poorly in each case. The results are presented in Table 17. All p-values from the Z-test are summarized in annex in Table 35.

Model	Forecast Period	Exogenous	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	3	cloudcover*	67.81	84.67	40.22	3.887	-2.887
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	3		75.65	88.67		-1.016	0.197
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	6		68.68	94.75	67.67	0.153	0.098
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	6		255.87	235.71		0.259	-0.717
SARIMAX(0, 1, 1)x(0, 1, 1, 12)	12		99.00	160.73	117.61	-0.725	1.310
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive	12		358.44	338.23		0.224	-0.711

TABLE 17. SARIMAX Results for Product E 100. For all time periods, the variable with the lowest AIC was tested, as no significant variables were identified.

5.3. XGBM

The final family of models tested was Extreme Gradient Boosting Machines (XGBM), chosen for their resilience to the limited amount of training data available. Using Gradient Boosting Machines shifts the forecasting approach to a regression framework, which removes direct dependencies on date order. To preserve temporal context, additional time-based features are created, including extracted values for month, year, and quarter from the date. Lagged features are also added to capture seasonality, including a 12-month lag and shorter lags ranging from 1 month to the number of significant peaks observed in the PACF for each product.

As in previous forecasts, Cross-Validation was applied to assess the models' generalization performance. Additionally, Cross-Validation was used during grid search to identify the optimal hyper-parameters for each model, by testing a set of values on: n_estimators, learning_rate, subsample and max_depth. Max depth was the only set of values varying between products, such that it starts from value 3, in intervals of 2, until about 75% of the number of features being used - accounting with time features - and as so, the respective set of values used can be analyzed in Table 18, while the other parameters can be found in Table 19.

Product	Columns Count	max_depth
Q 75	12	[3,5,7,9]
Q 100	10	[3,5,7]
E 75	12	[3,5,7,9]
Q 6	10	[3,5,7]
E 100	13	[3,5,7,9]

TABLE 18. Range of values tested for max_depth for each product

Variable	Values
n_estimators	[50, 100, 200, 300]
learning_rate	[0.1, 0.01, 0.001]
subsample	[0.5, 0.7, 1]

TABLE 19. List of common product parameters with values for GridSearch tuning

Starting with Product Q 75, the obtained parameters and their results can be examined in Table 20. Additionally, feature importance across different periods can be analyzed in Figure 24. The top four features identified included the same variables but were ranked differently for the 3-month period, namely: the data point from one year ago (lag_1y), month, year, and IPI, while quarter information had no impact on the model.

Forecast Period	Model Parameters	RMSE	RMSE (CV)	Average (Test)	TISP	BIAS
3	learning_rate: 0.1 max_depth: 3 n_estimators: 100 subsample: 0.5	972.75	1120.40	1708.00	0.525	-0.010
6	learning_rate: 0.1 max_depth: 3 n_estimators: 300 subsample: 0.5	1282.50	1155.68	2960.25	0.680	-0.107
12	learning_rate: 0.1 max_depth: 3 n_estimators: 100 subsample: 0.5	1373.98	1243.56	3364.63	0.716	-0.184

TABLE 20. XGBM Models Results for Product Q 75

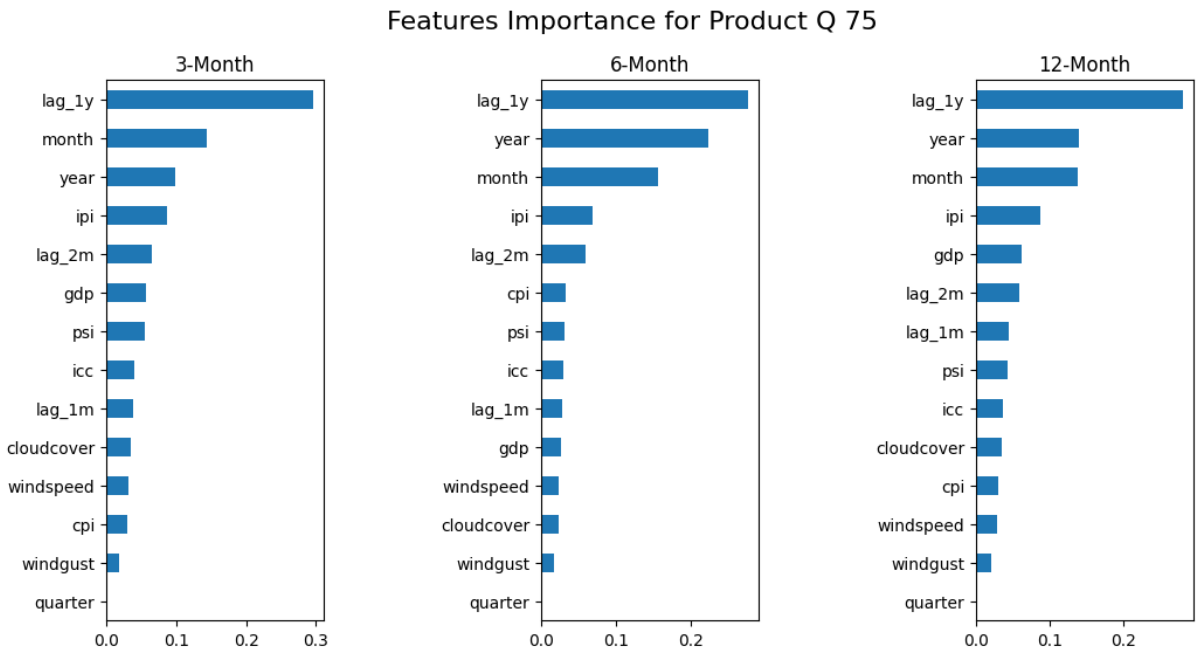


FIGURE 24. Features Importance for Product Q 75

For product Q 100, the obtained parameters and their results can be observed in Table 21. Additionally, feature importance across different periods can be analyzed in Figure 25. Value of the past year was the feature with most importance in all periods, and besides this variable, the 4 most important across periods, counted always with: year, month and the value of previous month. Quarter information, once again, showed no impact on the models.

Forecast Period	**Model Parameters**	**RMSE**	**RMSE (CV)**	**Average (Test)**	**TISP**	**BIAS**
3	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 0.7	459.41	707.50	197.56	0.313	-0.687
6	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 0.5	477.26	673.59	468.56	0.305	-0.085
12	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 0.7	377.17	639.33	547.22	0.447	0.167

TABLE 21. XGBM Models Results for Product Q 100

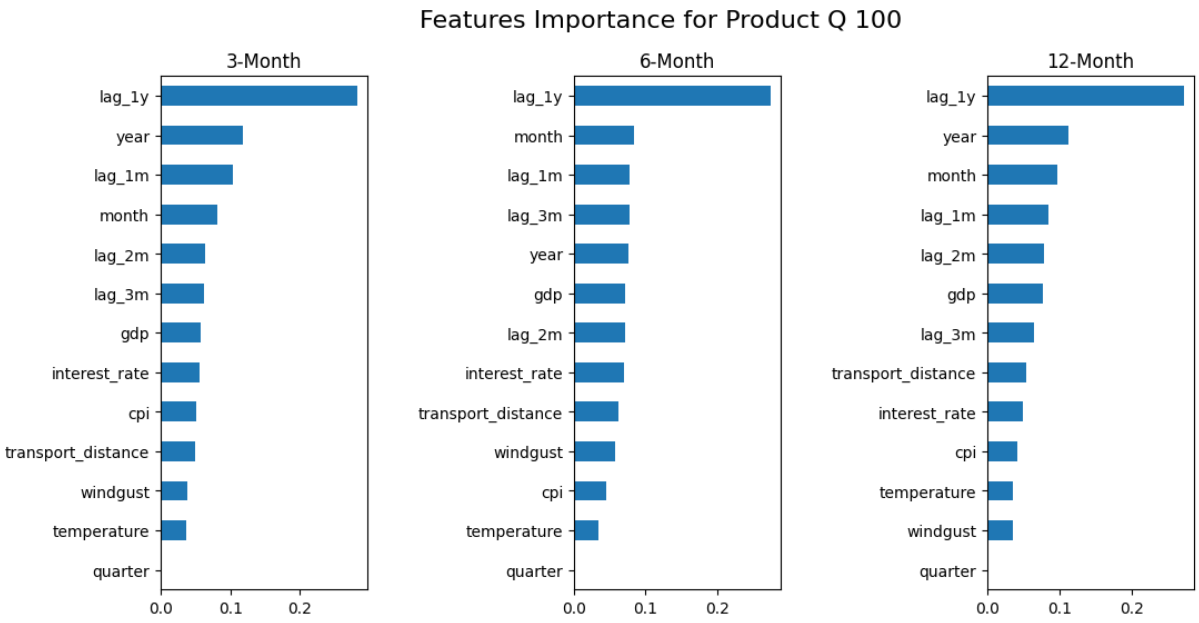


FIGURE 25. Features Importance for Product Q 100

For product E 75, the obtained parameters and their results can be examined in Table 22. Additionally, feature importance across different periods can be analyzed in Figure 26. Interestingly, for 6-month period, it was used a higher *max_depth* of 9, compared to previous models. The top 3 most influential features were the same through all the periods: the value of last year, year, and month, just varying the two leading variables between the first two. Quarter information kept showing no impact on the models.

Forecast Period	**Model Parameters**	**RMSE**	**RMSE (CV)**	**Average (Test)**	**TISP**	**BIAS**
3	learning_rate: 0.1 max_depth: 3 n_estimators: 300 subsample: 0.5	312.06	407.70	474.17	0.375	0.107
6	learning_rate: 0.1 max_depth: 9 n_estimators: 200 subsample: 0.5	347.61	408.54	823.00	0.693	-0.085
12	learning_rate: 0.1 max_depth: 7 n_estimators: 200 subsample: 0.7	445.16	406.25	968.12	0.639	-0.110

TABLE 22. XGBM Models Results for Product E 75

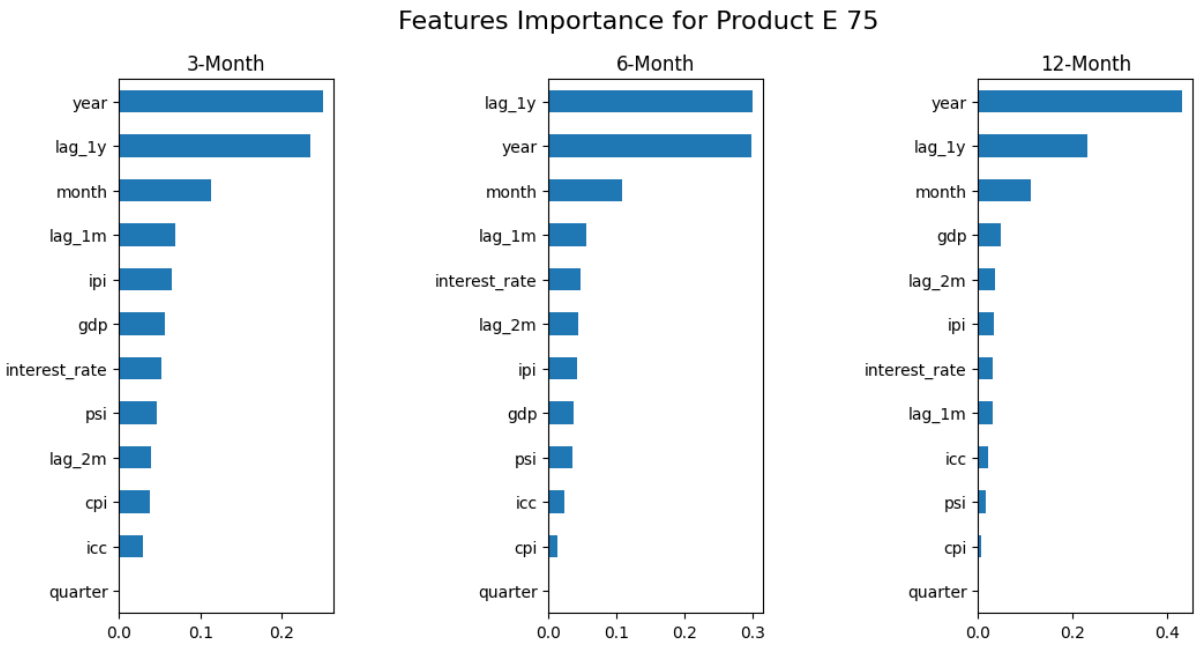


FIGURE 26. Features Importance for Product E 75

For product Q 6, the obtained parameters and their results can be examined in Table 23. Additionally, feature importance across different periods can be analyzed in Figure 27. The top 3 most influential features — year, value from last year, and month — remained consistent in both presence and ranking across all periods. Interestingly, the year variable had a low influence on this product, even becoming non-significant in the 12-month period. This contrasts with previous products, where year was the most influential feature, consistently topping the chart. Quarter information kept showing no impact on the models.

Forecast Period	**Model Parameters**	**RMSE**	**RMSE (CV)**	**Average (Test)**	**TISP**	**BIAS**
3	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 0.7	175.21	164.82	176.89	0.288	-0.261
6	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 0.5	142.20	156.86	220.40	0.545	-0.230
12	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 1	127.87	153.64	303.62	0.713	-0.161

TABLE 23. XGBM Models Results for Product Q 6

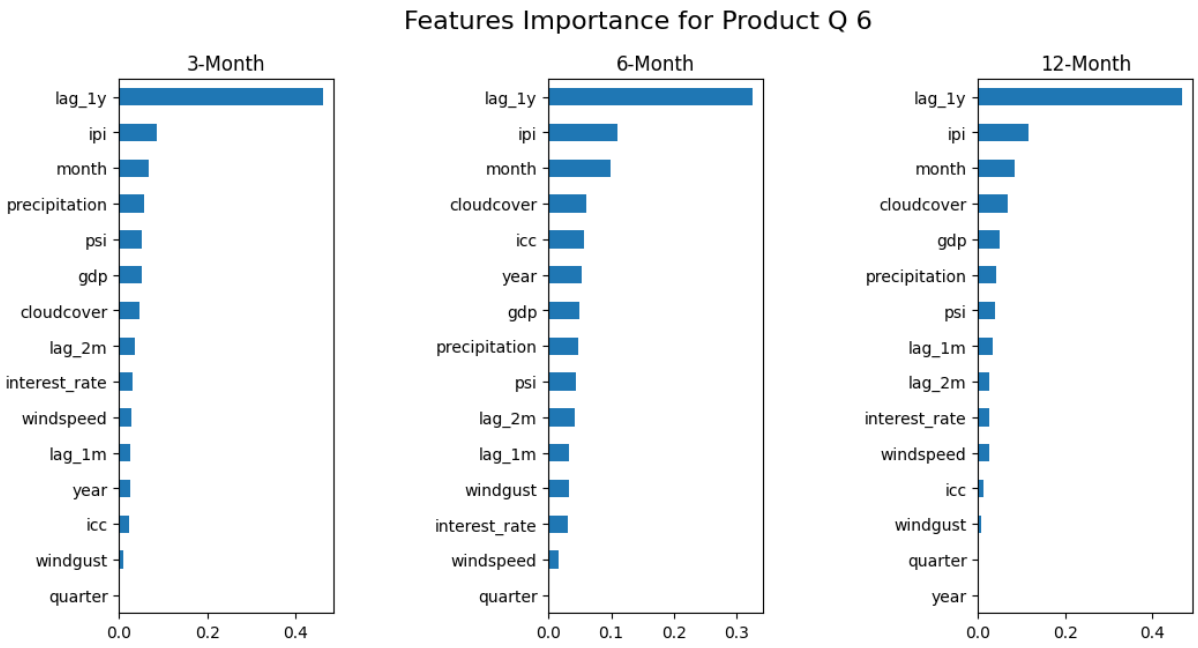


FIGURE 27. Features Importance for Product Q 6

For product E 100, the obtained parameters and their results can be examined in Table 24. Additionally, feature importance across different periods can be analyzed in Figure 28. The top 3 most influential features — value from last year, value from last month and month — remained consistent in both presence and ranking across all periods. Quarter information continued to show no impact on the models, and with this being the final product analysis, it proved to be a non-informative variable in every case.

Forecast Period	**Model Parameters**	**RMSE**	**RMSE (CV)**	**Average (Test)**	**TISP**	**BIAS**
3	learning_rate: 0.01 max_depth: 3 n_estimators: 300 subsample: 0.5	62.42	251.40	40.22	0.408	-0.592
6	learning_rate: 0.01 max_depth: 3 n_estimators: 200 subsample: 0.5	121.22	241.06	67.67	0.395	-0.605
12	learning_rate: 0.1 max_depth: 3 n_estimators: 50 subsample: 0.5	84.94	244.14	117.61	0.530	-0.097

TABLE 24. XGBM Models Results for Product E 100

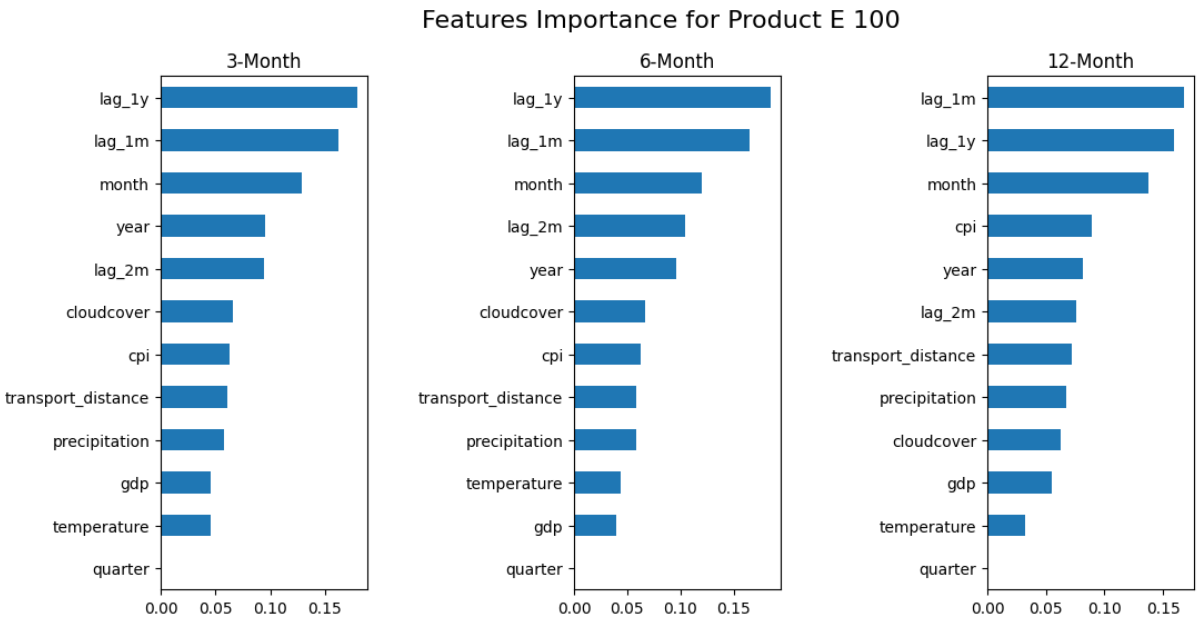


FIGURE 28. Features Importance for Product E 100

Following the predictions and the assessment of the most significant features for each model, the top five features from each model were compiled, and their occurrences were counted for analysis by period, as well as overall across all periods. As shown in Figure 29, the three most frequent features in each period remained consistent: month, value from the previous year, and year.

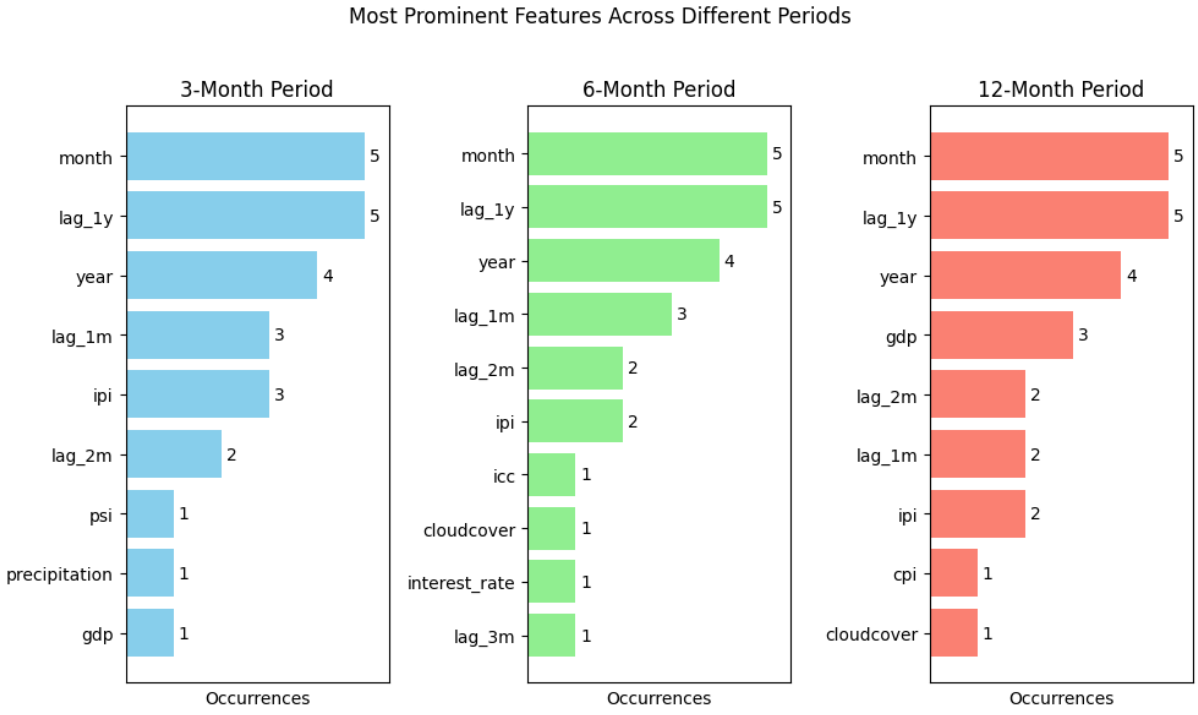


FIGURE 29. Most Important Features for Each Period

When examining the overall results across the three merged periods, in addition to the top three features previously mentioned, several other variables emerged with more than one occurrence. These include past values from one and two months prior, as well as IPI, GDP, and notably, cloud cover, as illustrated in Figure 30.

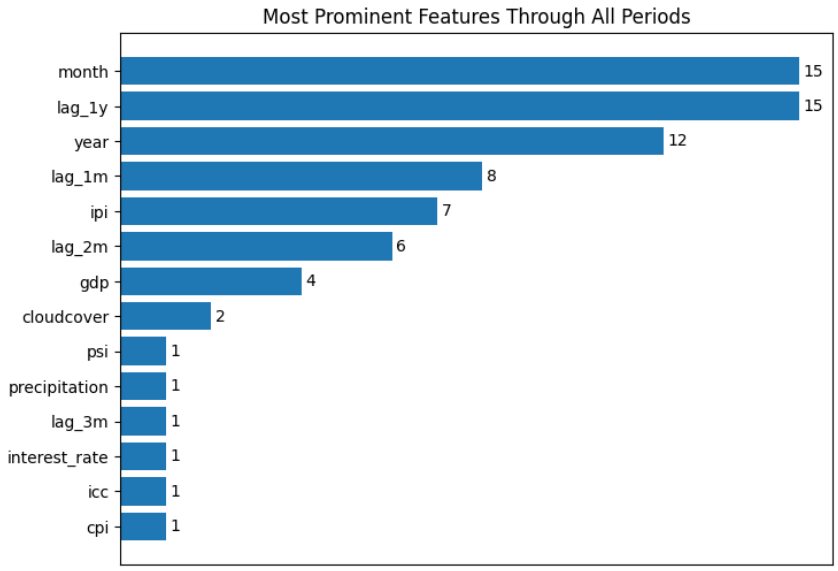


FIGURE 30. Most Important Features Through All Periods

5.4. Comparisons and Selections

After completing the baseline predictions with SARIMA, along with the tests using SARIMAX and Gradient Boosting Machines, we can now compare model performance across products and time periods. This comparison will help identify the best-performing model for each scenario and assess how SARIMAX and GBMs performed relative to the baseline. For each product, a table presents the best model achieved within each model type for each time period, and the best model for the product on a given period, is highlighted in yellow.

Beginning with Q 75, as shown in Table 25, SARIMAX performs best for the 3-month prediction, while XGBMs did not surpass the baseline. For the 6-month and 12-month predictions, the baseline model achieved the highest accuracy, followed by SARIMAX and then XGBMs. Taking in account the best models selected on the table for this product, only one model is using an exogenous variable, which is ICC on the 3-month period.

Model	Forecast Period	Average (Test)	RMSE (CV)
SARIMA(1,1,1)(0,1,2)[12]	3	1708.00	920.27
SARIMAX(1, 1, 1)x(0, 1, [1, 2], 12)			859.03
XGBM			1120.40
SARIMA(1,1,1)(1,1,1)[12]	6	2960.25	948.96
SARIMAX(1, 1, 1)x(1, 1, 1, 12)			1012.93
XGBM			1155.68
SARIMA(1,1,1)(1,1,1)[12] Recursive	12	3364.63	1002.26
SARIMAX(1, 1, 1)x(1, 1, 1, 12) Recursive			1207.09
XGBM			1243.56

TABLE 25. Comparison of Models Evaluation on Q 75

Moving to product Q 100, the table in Table 26 shows the same outcomes as the previous product, where the baseline has better results on both 6-month and 12-month periods, while for a 3-month period SARIMAX has the best result. XGBMs, once again had the worst performance in all cases. From the best-identified models, only the variable GDP is used in the 3-month period.

Model	Forecast Period	Average (Test)	RMSE (CV)
SARIMA(0,1,1)(0,1,1)[12] Recursive	3	197.56	442.06
SARIMAX(3, 1, 1)x(0, 1, 1, 12) Recursive			388.62
XGBM			707.50
SARIMA(1,1,1)(0,1,1)[12] Recursive	6	468.56	418.06
SARIMAX(3, 1, 1)x(0, 1, 1, 12)			426.71
XGBM			673.59
SARIMA(1,1,1)(0,1,1)[12] Recursive	12	547.22	480.94
SARIMAX(3, 1, 0)x(0, 1, [1], 12)			491.25
XGBM			639.33

TABLE 26. Comparison of Models Evaluation on Q 100

Examining a different bottle code through Table 27, reveals a distinct scenario: SARIMAX achieved the best results across all periods, while for the 12-month period, XGBMs performed better than the baseline. The top-performing models include the variables GDP, PSI, and ICC.

Model	Forecast Period	Average (Test)	RMSE (CV)
SARIMA(0,1,1)(0,1,1)[12]	3	474.17	362.73
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive			345.29
XGBM			407.70
SARIMA(0,1,1)(0,1,1)[12] Recursive	6	823.00	354.42
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive			321.31
XGBM			408.54
SARIMA(0,1,1)(0,1,1)[12]	12	968.12	406.80
SARIMAX(1, 1, 2)x(0, 1, [1], 12)			331.28
XGBM			406.25

TABLE 27. Comparison of Models Evaluation on E 75

Returning to group code Q, specifically Q 6, the same conclusions regarding the best model types were observed as with the other products of this bottle code. Looking at Table 28, SARIMAX outperformed the others in the 3-month period, while the baseline remained the top option for both the 6-month and 12-month periods. XGBMs, on the other hand, consistently exhibited the worst performance across all periods. The only variable used on best models, is PSI on the 3-month period.

Model	Forecast Period	Average (Test)	RMSE (CV)
SARIMA(2,1,0)(2,1,0)[12]	3	176.89	86.77
SARIMAX(0, 1, 1)x(0, 1, 1, 12) Recursive			73.78
XGBM			164.82
SARIMA(2,1,0)(0,1,1)[12] Recursive	6	220.40	106.54
SARIMAX(0, 1, 1)x(0, 1, 1, 12)			122.61
XGBM			156.86
SARIMA(2,1,0)(2,1,0)[12] Recursive	12	303.62	126.71
SARIMAX(2, 1, 1)x(2, 1, [], 12)			127.91
XGBM			153.64

TABLE 28. Comparison of Models Evaluation on Q 6

On E 100, as observed in Table 29, there's a similar pattern of best models as in products of group code Q, where SARIMAX is the best model for the 3-month period, using the variable

of cloud cover, while the baseline is the best model for the remaining periods. Once again, XGBMs have the worst performance in all the periods.

Model	Forecast Period	Average (Test)	RMSE (CV)
SARIMA(0,1,1)(0,1,1)[12] Recursive	3	40.22	90.77
SARIMAX(0, 1, 1)x(0, 1, 1, 12)			84.67
XGBM			251.40
SARIMA(0,1,1)(0,1,1)[12] Recursive	6	67.67	93.03
SARIMAX(0, 1, 1)x(0, 1, 1, 12)			94.75
XGBM			241.06
SARIMA(1,1,1)(0,1,1)[12] Recursive	12	117.61	144.50
SARIMAX(0, 1, 1)x(0, 1, 1, 12)			160.73
XGBM			244.14

TABLE 29. Comparison of Models Evaluation on E 100

In conclusion, XGBMs consistently had the lowest performance, while the best models alternated between SARIMA models with and without exogenous variables. The variables utilized in these top-performing models included GDP, PSI, ICC, and cloud cover.

CHAPTER 6

Conclusion

The primary objective of this study was to forecast the sales volume of spirit drinks in the off-trade channel of a company, over three different time periods—3, 6, and 12 months—while also identifying external factors that influence sales. The significance of this study lies not only in its potential business applications—enhancing forecasts beyond the current manual methods used by the company—but also in its contribution to the literature. By analyzing the sales of alcoholic beverages from a manufacturer's perspective within the off-trade channel, this research provides unique insights that are rarely explored in existing literature. Furthermore, it acknowledges the existence of intermediaries and the variability of retailer policies as indirect factors that may influence sales outcomes.

To accomplish this, an initial Exploratory Data Analysis was conducted to gain insights into the context of the variables, followed by a direct analysis of sales behavior for each product. For instance, the impact of COVID-19 was specifically examined. Only after gaining a better understanding of the data were the models developed, and potential exogenous variables were tested.

The resultant models varied between SARIMA and SARIMAX, and the exogenous variables observed were: GDP, PSI, ICC, and cloud cover.

6.1. Models Reflections

Three types of models were employed to analyze the sales data: SARIMA, SARIMAX, and XGBoost. In addition to discussing the performance of these models, this section reflects on various factors that may have contributed to results that deviated from expectations, as well as the prevailing conditions that influenced these outcomes.

Firstly, the dataset comprised a limited set of 132 observations for each product, which poses significant challenges for a deep learning model to effectively capture patterns. Such models are typically recommended in the literature for their ability to yield better results and detect non-linear patterns. This limitation may explain the inferior performance of the XGB models; however, it is noteworthy that in some instances, their results were relatively close to those of the other models (as referenced in section 5.4).

Regarding the exogenous variables, temperature was anticipated to significantly influence sales, as discussed in the dedicated section on weather in the literature review (2.3.1), which highlighted its effects. I suggest two potential reasons for the observed lack of impact in this study: first, the weather dataset is limited to the city of Lisbon, which may not accurately reflect weather patterns across the entire country; second, the use of monthly data aggregates temperature readings, potentially obscuring any nuanced effects that might manifest on a more

granular level. Consequently, the influence of temperature may not be evident when analyzed at a monthly scale, as typically literature studies were done with daily or weekly data.

In terms of sales data, it was not made an outlier analysis, which would be important, especially because of Covid, which had a negative effect on two products, as was seen in subsection 4.4.1.

6.2. Future Work

While this study did an extensive search, especially on the literature review and exploratory data analysis part, there are still factors that could be done to further improve the model's results and capture variables affecting the sales; some aspects come from the previous section.

Given no outlier analysis was done, this could be the start point. In terms of products, it was only analyzed the top 5 most sold, so there are more products to be analyzed, and an original thought I had was to create clusters with products to help with the predictions. In terms of sales data, it would be useful to have a more granular data, in terms of time, but that is not possible, nonetheless, data from marketing events could help the predictions.

From exogenous variables, it could be tried to summarize national weather, and tourism is also a field to explore, as it was a factor mentioned at the board meeting.

References

Abolghasemi, M., Hurley, J., Eshragh, A., & Fahimnia, B. (2020). Demand forecasting in the presence of systematic events: Cases in capturing sales promotions. *International Journal of Production Economics*, *230*, 107892. https://doi.org/https://doi.org/10.1016/j.ijpe.2020.107892

Agnew, M., & Thornes, J. (2007). The weather sensitivity of the uk food retail and distribution industry. *Meteorological Applications*, *2*, 137–147. https://doi.org/10.1002/met.5060020207

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales:: A comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, *8*(3), 147–156. https://doi.org/https://doi.org/10.1016/S0969-6989(00)00011-4

Arunraj, N. S., & Ahrens, D. (2016). Estimation of non-catastrophic weather impacts for retail industry [Cited by: 25]. *International Journal of Retail and Distribution Management*, *44*(7), 731–753. https://doi.org/10.1108/IJRDM-07-2015-0101

Babai, M., Ali, M., Boylan, J., & Syntetos, A. (2013). Forecasting and inventory performance in a two-stage supply chain with arima(0,1,1) demand: Theory and empirical analysis [Focusing on Inventories: Research and Applications]. *International Journal of Production Economics*, *143*(2), 463–471. https://doi.org/https://doi.org/10.1016/j.ijpe.2011.09.004

Badorf, F., & Hoberg, K. (2020). The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. *Journal of Retailing and Consumer Services*, *52*, 101921. https://doi.org/https://doi.org/10.1016/j.jretconser.2019.101921

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In T. Gedeon, K. W. Wong, & M. Lee (Eds.), *Neural information processing* (pp. 462–474). Springer International Publishing.

Bauerová, R., Starzyczná, H., & Pražák, T. (2022). At what lag should economic indicators be applied to predict sales in a transformed economy? is it worthwhile when e-tailing? [Cited by: 0]. *Forum Scientiae Oeconomia*, *10*(4), 29–46. https://doi.org/10.23762/FSO_VOL10_NO4_2

Bujisic, M., Bogicevic, V., Parsa, H. G., Jovanovic, V., & Sukhu, A. (2019). It's raining complaints! how weather factors drive consumer comments and word-of-mouth. *Journal of Hospitality & Tourism Research*, *43*(5), 656–681. https://doi.org/10.1177/1096348019835600

Chan, H., & Wahab, M. (2024). A machine learning framework for predicting weather impact on retail sales. *Supply Chain Analytics*, *5*, 100058. https://doi.org/https://doi.org/10.1016/j.sca.2024.100058

Chawla, A., Singh, A., Lamba, A., Gangwani, N., & Soni, U. (2019, September). Demand forecasting using artificial neural networks—a case study of american retail corporation. https://doi.org/10.1007/978-981-13-1822-1_8

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly (1970-1977)*, *23*(3), 289–303. Retrieved January 4, 2024, from http://www.jstor.org/stable/3007885

Eaves, A. H. C., & Kingsman, B. G. (2004). Forecasting for the ordering and stock-holding of spare parts. *Journal of the Operational Research Society*, *55*(4), 431–437. https://doi.org/10.1057/palgrave.jors.2601697

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice [Special Issue: M5 competition]. *International Journal of Forecasting*, *38*(4), 1283–1318. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.06.004

Gao, J., Xie, Y., Cui, X., Yu, H., & Gu, F. (2018). Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model. *Advances in Mechanical Engineering*, *10*, 168781401774932. https://doi.org/10.1177/1687814017749325

Gardner, M. P., & Hill, R. P. (1988). Consumers' mood states: Antecedents and consequences of experiential versus informational strategies for brand choice [Cited by: 24]. *Psychology & Marketing*, *5*(2), 169–182. https://doi.org/10.1002/mar.4220050206

Gaur, M., Goel, S., & Jain, E. (2015). Comparison between nearest neighbours and bayesian network for demand forecasting in supply chain management. *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1433–1436.

Ghobbar, A. A., & Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model. *Computers & Operations Research*, *30*(14), 2097–2114. https://doi.org/https://doi.org/10.1016/S0305-0548(02)00125-9

Groene, N., & Zakharov, S. (2024). Introduction of ai-based sales forecasting: How to drive digital transformation in food and beverage outlets. *Discover Artificial Intelligence*, *4*. https://doi.org/10.1007/s44163-023-00097-x

Gustriansyah, R., Suhandi, N., Antony, F., & Sanmorino, A. (2019). Single exponential smoothing method to predict sales multiple products. *Journal of Physics: Conference Series*, *1175*(1), 012036. https://doi.org/10.1088/1742-6596/1175/1/012036

Hirche, M., Haensch, J., & Lockshin, L. (2021). Comparing the day temperature and holiday effects on retail sales of alcoholic beverages – a time-series analysis. *International Journal of Wine Business Research*, *ahead-of-print*. https://doi.org/10.1108/IJWBR-07-2020-0035

Hu, J., Zhang, X., Chen, H., & Li, W. (2024). When it rains, it pours? the impact of weather on customer returns in the brick-and-mortar retail store. *Journal of Retailing and Consumer Services*, *77*, 103664. https://doi.org/https://doi.org/10.1016/j.jretconser.2023.103664

KALAOGLU, O. I., AKYUZ, E. S., ECEMİŞ, S., ERYURUK, S. H., SÜMEN, H., & KALAOGLU, F. (2015). Retail demand forecasting in clothing industry. *Textile and Apparel*, *25*(2), 172–178.

Kapoor, M., & Ravi, S. (2009). The effect of interest rate on household consumption: Evidence from a natural experiment in india. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1346813

Keleş, B., Gómez-Acevedo, P., & Shaikh, N. I. (2018). The impact of systematic changes in weather on the supply and demand of beverages. *International Journal of Production Economics*, *195*, 186–197. https://doi.org/https://doi.org/10.1016/j.ijpe.2017.08.002

Keller, M. C., Fredrickson, B. L., Ybarra, O., Côté, S., Johnson, K., Mikels, J., Conway, A., & Wager, T. (2005). A warm heart and a clear head: The contingent effects of weather on mood and cognition [Cited by: 300; All Open Access, Green Open Access]. *Psychological Science*, *16*(9), 724–731. https://doi.org/10.1111/j.1467-9280.2005.01602.x

Khajehzadeh, M., Pazhuheian, F., Seifi, F., Noorossana, R., Asli, S., & Saeedi, N. (2022). Analysis of factors affecting product sales with an outlook toward sale forecasting in cosmetic industry using statistical methods. *International Review of Management and Marketing*, *12*, 55–63. https://doi.org/10.32479/irmm.13337

Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on dwt decomposition [International Conference on Computer, Communication and Convergence (ICCC 2015)]. *Procedia Computer Science*, *48*, 173–179. https://doi.org/https://doi.org/10.1016/j.procs.2015.04.167

Krichene, E., Masmoudi, Y., Alimi, A. M., Abraham, A., & Chabchoub, H. (2017). Forecasting using elman recurrent neural network. In A. M. Madureira, A. Abraham, D. Gamboa, & P. Novais (Eds.), *Intelligent systems design and applications* (pp. 488–497). Springer International Publishing.

Kuo, R. J., Tseng, Y. S., & Chen, Z.-Y. (2016). Integration of fuzzy neural network and artificial immune system-based back-propagation neural network for sales forecasting using qualitative and quantitative data. *Journal of Intelligent Manufacturing*, *27*(6), 1191–1207. https://doi.org/10.1007/s10845-014-0944-1

Kuo, R. (2001). A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research*, *129*(3), 496–517. https://doi.org/https://doi.org/10.1016/S0377-2217(99)00463-4

Lazo, J. K., Lawson, M., Larsen, P. H., & Waldman, D. M. (2011). U.s. economic sensitivity to weather variability [Cited by: 108]. *Bulletin of the American Meteorological Society*, *92*(6), 709–720. https://doi.org/10.1175/2011BAMS2928.1

Lin, C.-J., & Lee, T.-S. (2013). Tourism Demand Forecasting: Econometric Model based on Multivariate Adaptive Regression Splines, Artificial Neural Network and Support Vector

Regression. *Advances in Management and Applied Economics*, *3*(6), 1–1. https://ideas.repec.org/a/spt/admaec/v3y2013i6f3_6_1.html

Liu, N., Ren, S., Choi, T.-M., Hui, C.-L., & Ng, S.-F. (2013). Sales Forecasting for Fashion Retailing Service Industry: A Review (K. Govindan, Ed.). *Mathematical Problems in Engineering*, *2013*, 738675. https://doi.org/10.1155/2013/738675

Loureiro, A., Miguéis, V., & da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81–93. https://doi.org/https://doi.org/10.1016/j.dss.2018.08.010

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111–128. https://doi.org/https://doi.org/10.1016/j.ejor.2020.05.038

Martínez-de-Albéniz, V., & Belkaid, A. (2021). Here comes the sun: Fashion goods retailing under weather fluctuations. *European Journal of Operational Research*, *294*(3), 820–830. https://doi.org/https://doi.org/10.1016/j.ejor.2020.01.064

Miranda-Moreno, L. F., & Lahti, A. C. (2013). Temporal trends and the effect of weather on pedestrian volumes: A case study of montreal, canada. *Transportation Research Part D: Transport and Environment*, *22*, 54–59. https://doi.org/https://doi.org/10.1016/j.trd.2013.02.008

Moon, S., Kang, M. Y., Bae, Y. H., & Bodkin, C. D. (2018). Weather sensitivity analysis on grocery shopping. *International Journal of Market Research*, *60*(4), 380–393. https://doi.org/10.1177/1470785317751614

Murray, K. B., Di Muro, F., Finn, A., & Popkowski Leszczyc, P. (2010). The effect of weather on consumer spending. *Journal of Retailing and Consumer Services*, *17*(6), 512–520. https://doi.org/https://doi.org/10.1016/j.jretconser.2010.08.006

Parnaudeau, M., & Bertrand, J.-L. (2018). The contribution of weather variability to economic sectors. *Applied Economics*, *50*(43), 4632–4649. https://doi.org/10.1080/00036846.2018.1458200

Pongdatu, G., & Putra, Y. Seasonal time series forecasting using sarima and holt winter's exponential smoothing. In: *407*. (1). 2018. https://doi.org/10.1088/1757-899X/407/1/012153

Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J. K., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, *58*(16), 4964–4979. https://doi.org/10.1080/00207543.2020.1735666

Ramkumar, G., & Srinivasan, D. C. (2020). Goods and services tax and consumer buying behaviour -a study. *PalArch's Journal of Archaeology of Egypt/ Egyptology*, *17*, 2777–2787.

Sadaei, H. J., de Lima e Silva, P. C., Guimarães, F. G., & Lee, M. H. (2019). Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series. *Energy*, *175*, 365–377. https://doi.org/https://doi.org/10.1016/j.energy.2019.03.081

Sanders, J. L., & Brizzolara, M. S. (1982). Relationships between weather and mood [PMID: 28143374]. *The Journal of General Psychology*, *107*(1), 155–156. https://doi.org/10.1080/00221309.1982.9709917

Shrieenidhi, A., Ruba Roshini, S., & Ranjana, P. Walmart sales forecasting using time series analysis. In: 2024. https://doi.org/10.1109/ADICS58448.2024.10533547

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting [M4 Competition]. *International Journal of Forecasting*, *36*(1), 75–85. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.03.017

Spiliotis, E., Makridakis, S., Semenoglou, A.-A., & Assimakopoulos, V. (2022). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research*, *22*(3), 3037–3061. https://doi.org/10.1007/s12351-020-00605-2

Steele, A. T. (1951). Weather's effect on the sales of a department store. *Journal of Marketing*, *15*(4), 436–443. https://doi.org/10.1177/002224295101500404

Steinker, S., Hoberg, K., & Thonemann, U. W. (2017). The value of weather information for e-commerce operations. *Production and Operations Management*, *26*(10), 1854–1874. https://doi.org/10.1111/poms.12721

Štulec, I., Petljak, K., & Naletina, D. (2019). Weather impact on retail sales: How can weather derivatives help with adverse weather deviations? *Journal of Retailing and Consumer Services*, *49*, 1–10. https://doi.org/https://doi.org/10.1016/j.jretconser.2019.02.025

Syntetos, A., & Boylan, J. (2001). On the bias of intermittent demand estimates [Tenth International Symposium on Inventories]. *International Journal of Production Economics*, *71*(1), 457–466. https://doi.org/https://doi.org/10.1016/S0925-5273(00)00143-2

Syntetos, A., & Boylan, J. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*, 303–314. https://doi.org/10.1016/j.ijforecast.2004.10.001

Thiesing, F., & Vornberger, O. (1997). Sales forecasting using neural networks. *Proceedings of International Conference on Neural Networks (ICNN'97)*, *4*, 2125–2128 vol.4. https://doi.org/10.1109/ICNN.1997.614234

Tian, X., Cao, S., & Song, Y. (2021). The impact of weather on consumer behavior and retail performance: Evidence from a convenience store chain in china [Cited by: 29]. *Journal of Retailing and Consumer Services*, *62*. https://doi.org/10.1016/j.jretconser.2021.102583

Wang, C.-H. (2022). Considering economic indicators and dynamic channel interactions to conduct sales forecasting for retail sectors [Cited by: 20]. *Computers and Industrial Engineering*, *165*. https://doi.org/10.1016/j.cie.2022.107965

Wang, C.-H., & Gu, Y.-W. (2022). Sales forecasting, market analysis, and performance assessment for us retail firms: A business analytics perspective [Cited by: 2; All Open Access, Gold Open Access]. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178480

Wang, S., Huang, C.-T., Wang, W.-L., & Chen, Y.-H. (2010). Incorporating arima forecasting and service-level based replenishment in rfid-enabled supply chain. *International Journal of Production Research*, *48*(9), 2655–2677. https://doi.org/10.1080/00207540903564983

Weng, T., Liu, W., & Xiao, J. (2020). Supply chain sales forecasting based on lightGBM and LSTM combination model. *Industrial Management & Data Systems*, *120*(2), 265–279. https://doi.org/10.1108/IMDS-03-2019-0170

Willemain, T. R., Smart, C. N., Shockor, J. H., & DeSautels, P. A. (1994). Forecasting intermittent demand in manufacturing: A comparative evaluation of croston's method. *International Journal of Forecasting*, *10*(4), 529–538. https://doi.org/https://doi.org/10.1016/0169-2070(94)90021-3

Zhang, M., Huang, X.-n., & Yang, C.-b. (2020). A sales forecasting model for the consumer goods with holiday effects. *Journal of Risk Analysis and Crisis Response*, *10*, 69–76. https://doi.org/10.2991/jracr.k.200709.001

Zwebner, Y., Lee, L., & Goldenberg, J. (2014). The temperature premium: Warm temperatures increase product valuation [Cited by: 93]. *Journal of Consumer Psychology*, *24*(2), 251–259. https://doi.org/10.1016/j.jcps.2013.11.003

APPENDIX A

Additional Data Tables

SKU	Count	Mean	Standard Deviation	Minimum	Q1	Q2	Q3	Maximum
Q 75	132	2930.947	1524.546	341.500	1786.250	2810.000	3837.000	6797.500
Q 100	132	1224.397	899.455	10.667	542.833	947.333	1905.333	3726.667
E 75	132	861.606	483.154	39.000	499.875	797.250	1181.250	1980.000
Q 6	132	480.399	233.670	19.733	319.267	439.333	605.067	1285.133
E 100	132	317.168	283.786	4.667	106.167	207.667	438.333	1146.667

TABLE 30. Statistical Measures of Target Products

APPENDIX B

Data Distribution on Histograms

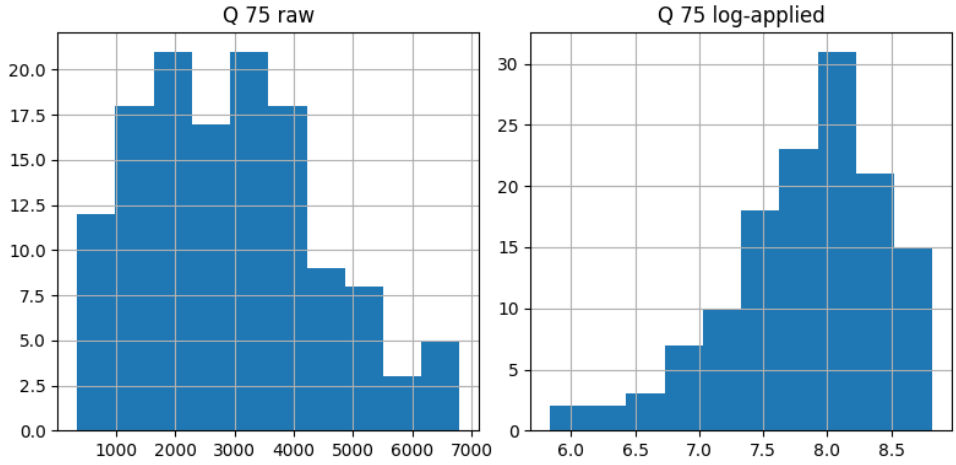


FIGURE 31. Data Distribution of Product Q 75 before and after applying logarithm

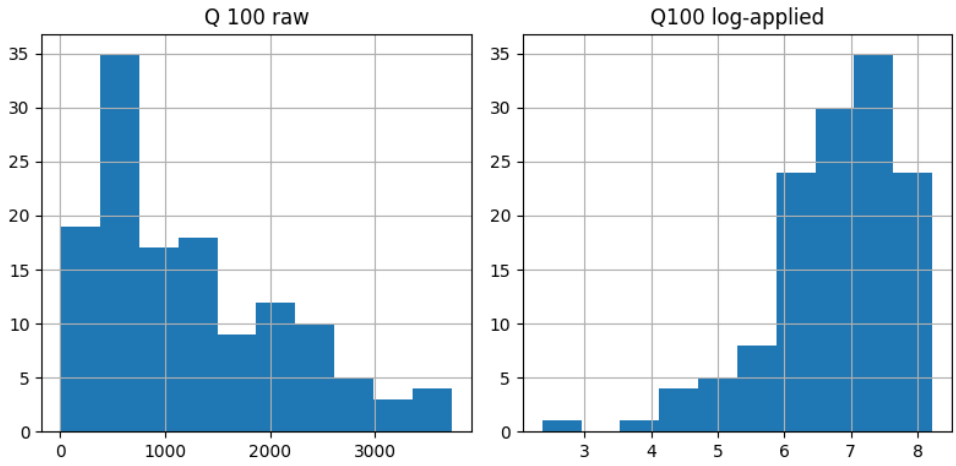


FIGURE 32. Data Distribution of Product Q 100 before and after applying logarithm

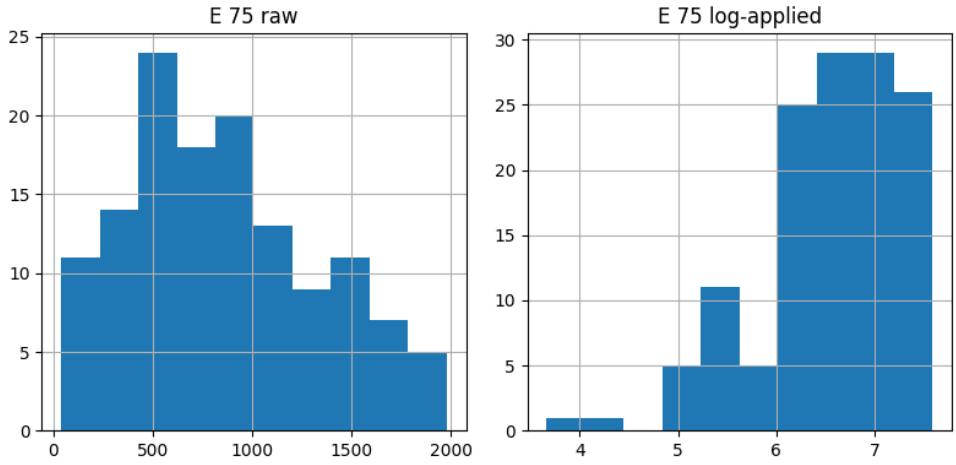


FIGURE 33. Data Distribution of Product E 75 before and after applying logarithm

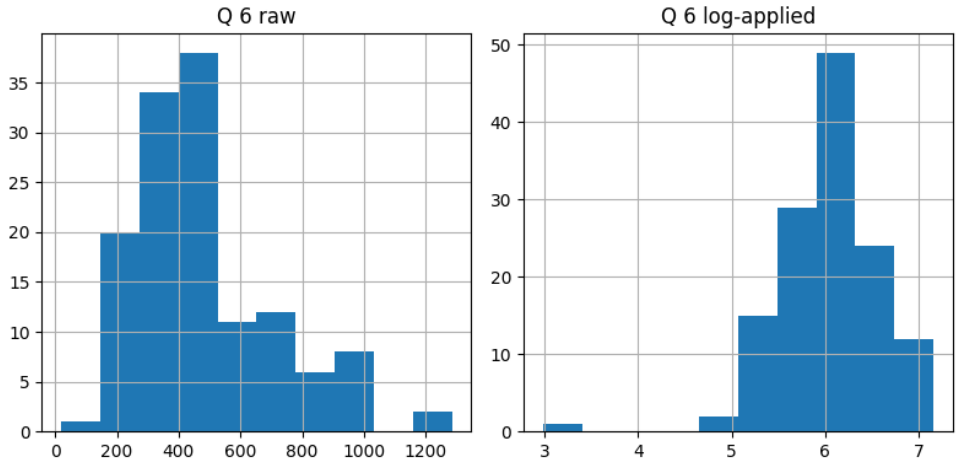


FIGURE 34. Data Distribution of Product Q 6 before and after applying logarithm

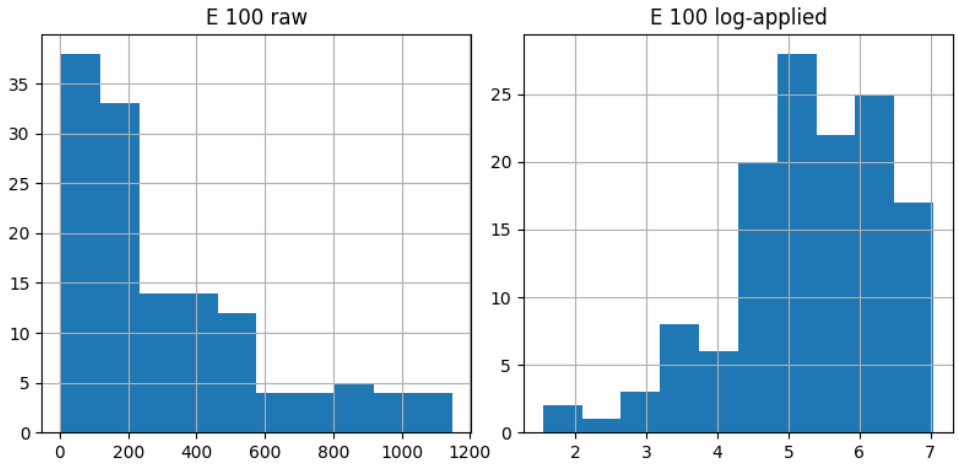


FIGURE 35. Data Distribution of Product E 100 before and after applying logarithm

APPENDIX C

Series Decompositions

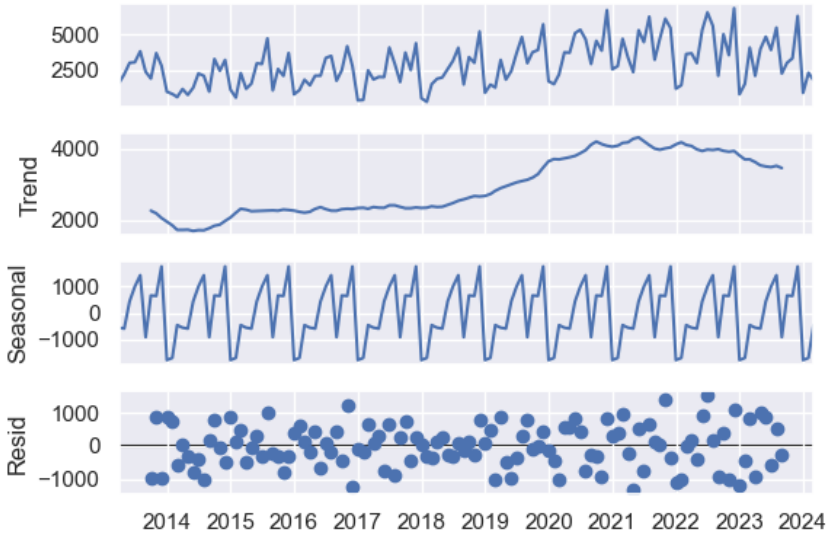


FIGURE 36. Decomposition of product Q 75

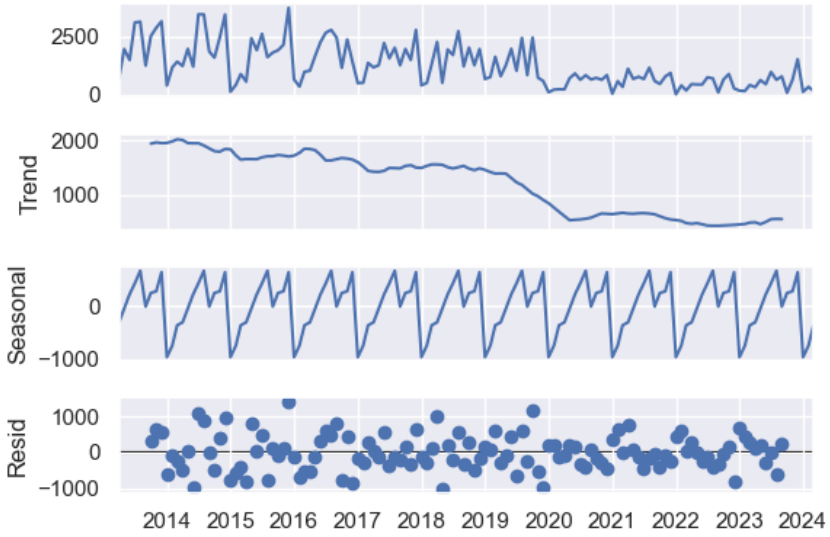


FIGURE 37. Decomposition of product Q 100

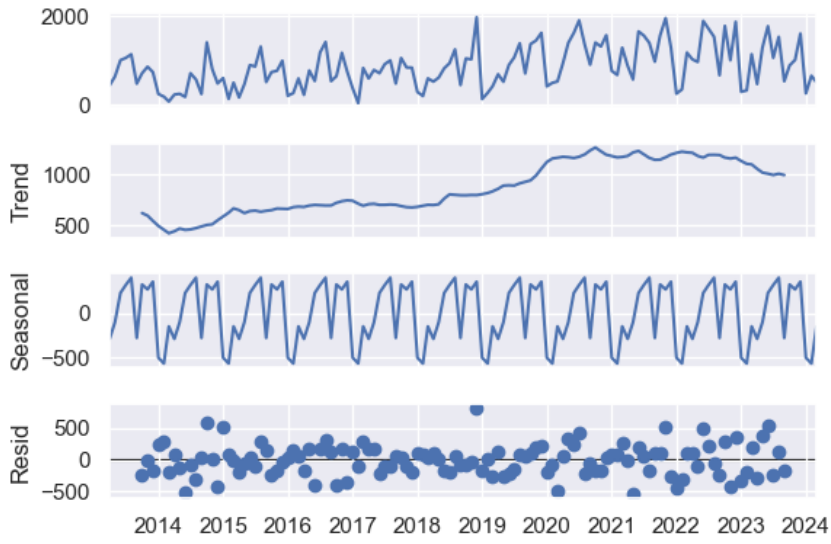


FIGURE 38. Decomposition of product E 75

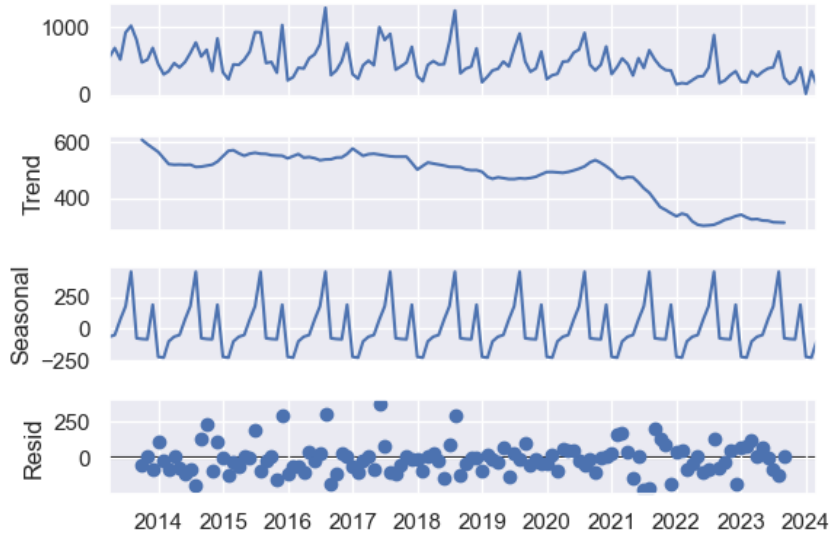


FIGURE 39. Decomposition of product Q 6

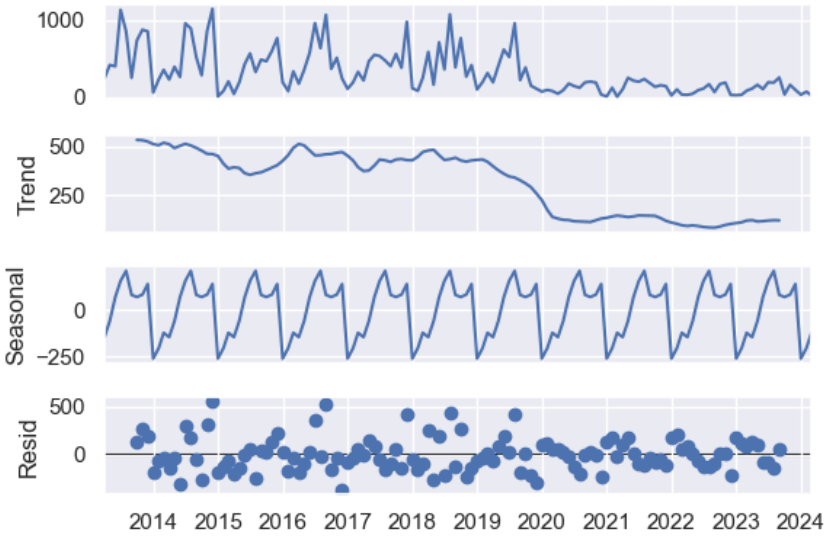


FIGURE 40. Decomposition of product E 100

APPENDIX D

Autocorrelation and Parcial Autocorrelation Analysis

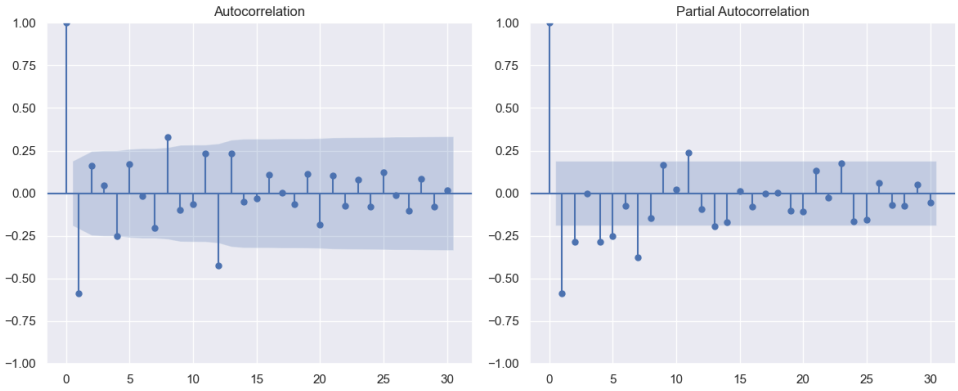


FIGURE 41. Autocorrelation and Partial Autocorrelation on Q 75 time series

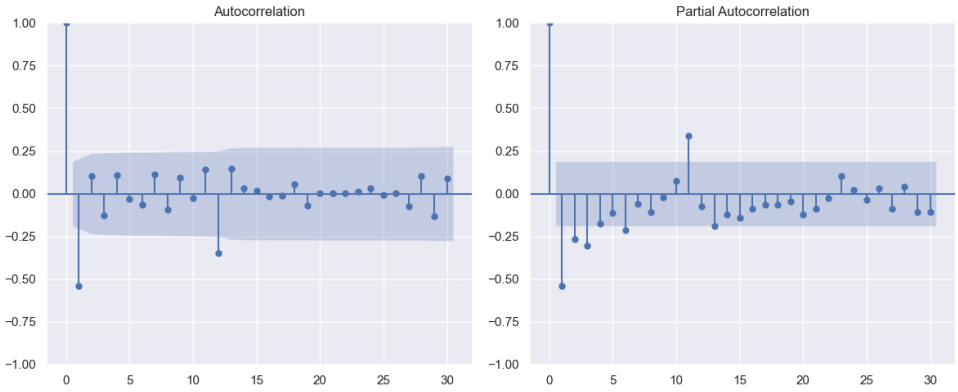


FIGURE 42. Autocorrelation and Partial Autocorrelation on Q 100 time series

FIGURE 43. Autocorrelation and Partial Autocorrelation on E 75 time series

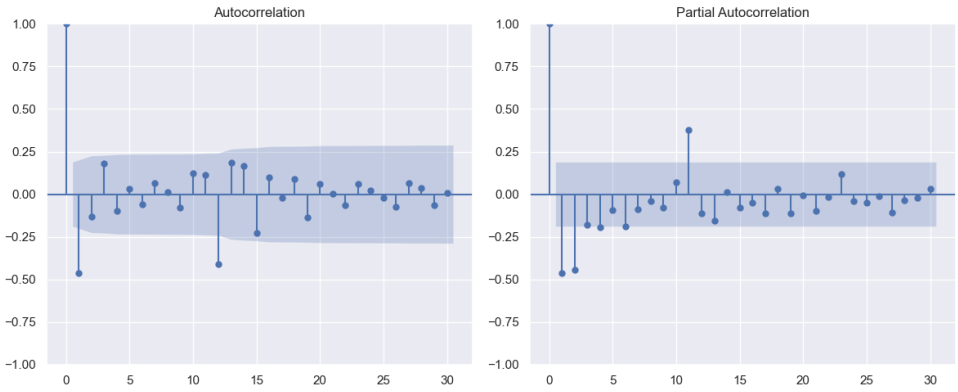


FIGURE 44. Autocorrelation and Partial Autocorrelation on Q 6 time series

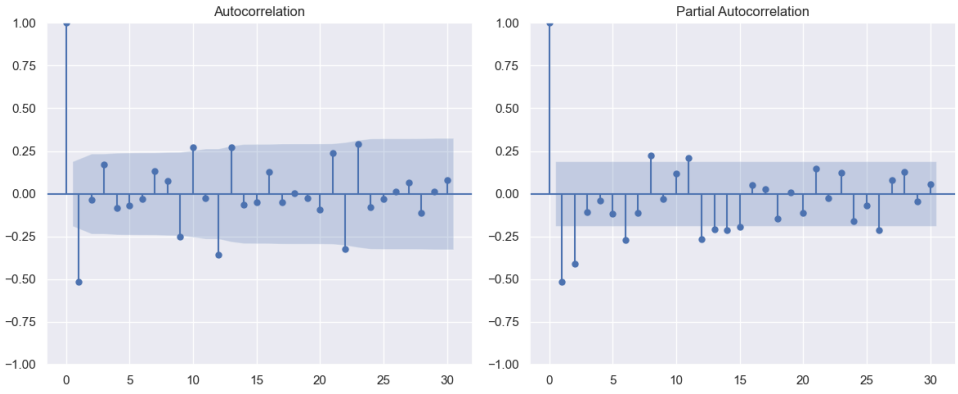


FIGURE 45. Autocorrelation and Partial Autocorrelation on E 100 time series

APPENDIX E

SARIMAX

variable	3M	6M	12M
windgust	0.317	0.226	0.308
windspeed	0.292	0.204	0.304
cloudcover	0.999	0.817	0.880
cpi	0.569	0.721	0.175
icc	0.064	0.066	0.109
gdp	0.396	0.173	0.056
ipi	0.832	0.984	0.614
psi	0.363	0.366	0.396

TABLE 31. Z-Test Results for Product Q 75 Models with Individual Exogenous Variables Across Time Periods

variable	3M	6M	12M
temperature	0.780	0.540	0.023
windgust	0.484	0.538	0.643
cpi	0.646	0.972	0.348
gdp	0.002	0.001	0.504
interest_rate	0.909	0.941	0.602
transport_distance	0.330	0.146	0.312

TABLE 32. Z-Test Results for Product Q 100 Models with Individual Exogenous Variables Across Time Periods

variable	3M	6M	12M
cpi	0.344	0.495	0.498
icc	0.163	0.030	0.001
gdp	0.002	0.913	0.958
ipi	0.352	0.341	0.505
interest_rate	0.504	0.577	0.607
psi	0.017	0.013	0.008

TABLE 33. Z-Test Results for Product E 75 Models with Individual Exogenous Variables Across Time Periods

variable	3M	6M	12M
precipitation	0.404	0.470	0.463
windgust	0.996	0.433	0.944
windspeed	0.593	0.233	0.243
cloudcover	0.750	0.823	0.777
icc	0.972	0.976	0.843
gdp	0.365	0.392	0.201
ipi	0.990	0.993	0.966
interest_rate	0.703	0.749	0.736
psi	0.071	0.863	0.084

TABLE 34. Z-Test Results for Product Q 6 Models with Individual Exogenous Variables Across Time Periods

variable	3M	6M	12M
temperature	0.934	0.972	0.970
precipitation	0.996	0.915	0.955
cloudcover	0.290	0.302	0.350
cpi	0.889	0.834	0.881
gdp	0.943	0.964	0.976
transport_distance	0.734	0.662	0.438

TABLE 35. Z-Test Results for Product E 100 Models with Individual Exogenous Variables Across Time Periods