NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Siemens Advanta Time Series Forecasting

Duarte Mendes, number: 20230494

Dzmitry Nisht, number: 20230776

Inês Silva, number: r20201580

José Marçal, number: r20201581

Ricardo Sousa, number: r20201611

Group E

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

1. EXECUTIVE SUMMARY

Siemens is a global company involved in a wide range of industries providing a great diversity of services. The filial Siemens Advanta understands the need to be updated on market trends and sales patterns to have a competitive advantage. Consequently, in this report, it shall be explored the process of developing forecasting models that allow to understand sales by group and the macroeconomic factors that influence them.

CRISP-DM was the methodology used to guide through this project. Primarily, it was crucial to understand the business which includes looking into its objects, what are the success criteria that shall be used as baselines to assess how the business currently is doing and how can be improved. Subsequently, it was required to analyse the resources made available to accomplish the optimum goal and what issues the data has. Therefore, it is necessary to find solutions to deal with the problems and how to prepare it to fit it into the models.

Following this, models were created and evaluated so that some changed could be made in order to reach the one with the best quality possible. After satisfied with the results obtained, the final outcomes were understood and studied to develop deployment and maintenance plans that will allow Siemens to put them in practise.

To sum up, this document explains the process to reach the final project that will facilitate the decision-making process while gaining market competitive advantage.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

This report follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) so in this part, it is possible to understand the organisation.

2.1. BUSINESS INTRODUCTION

Siemens Advanta, specifically, the consulting division, fosters sustainable and digital tailor-made solutions to fit into their customer's needs. Moreover, from all the characteristics that make this company outstanding, the one that gives them the biggest competitive advantage is the fact that they are equipped to work on a diversity of industries making them suitable to accept and compete for projects all over the world.

2.2. BUSINESS OBJECTIVES

The main purpose for this project consists in developing a precise sales prediction model for Siemens, capable of enabling effective sales and product patterns' comprehension over time, by applying advanced forecasting methods. In this way, we will be able to effectively assist Siemens by addressing critical business objectives, including identifying and analysing sales trends and discrepancies across different timeframes (yearly and monthly), or differentiating different groups of products over sales patterns, to enable targeted business strategies. Other business objectives are seen as fundamental for the company's business strategy, namely recommendations over sales margin improvement and increase of data transparency, based on an extensive analysis of the key factors affecting Siemens' Smart Infrastructure Division, aligning our efforts with the company's main goals.

In accordance, one of the anticipated key advantages for Siemens relates to the improvement of sales forecasting accuracy. The use of a combination of advanced forecasting and machine learning approaches and techniques enables better understanding of product sales, leading to more reliable sales predictions for the upcoming future.

Furthermore, the increase of sales margin will also appear as a related advantage, as this project's predictive insights and recommendations can ultimately lead to the optimization of Siemens' sales strategies and the overall profitability of the division and, subsequently, of the company as a whole.

As previously mentioned, data transparency is fundamental to provide clear insights and visibility into sales and product trends, allowing stakeholders to make informed decisions. Using the developed model, Siemens will be able to benefit from an improvement in the data transparency domain, integrating higher-quality insights and product and sales patterns that can serve the business strategy of Siemens.

Moreover, this project will also bring benefits to Siemens in terms of effective decision-making, in the sense that, by developing an accurate sales prediction model, Siemens will receive data-driven insights with the power to guide sustained decisions in diverse areas, such as production, marketing approaches or inventory management.

Additionally, the developed model will also provide data-driven and more effective pricing suggestions, ensuring competitive prices in the market, while maintaining optimal sales margins guided by Siemens' overall benefit.

Lastly, this project will also capacitate Siemens to provide higher-quality service to their customers, in terms of product availability and understanding of customer behaviour. A comprehensive understanding of sales and product trends at specific seasons of the year can lead to better meeting customer demand and expectations, as well as ensure appropriate inventory levels and, subsequently, enhance customer satisfaction and retention.

2.3. BUSINESS SUCCESS CRITERIA

A great part of the success of such project comes from establishing tangible and realistic business criteria, with capacity to properly assess its progress and respective outputs, and to evaluate the success of its implementation.

The identified business success criteria identified are mostly connected to the previously defined business objectives, such as the improvement in sales forecast accuracy of the model and reduction in forecast error, measured using Root Mean Square Error (RMSE) metric. For both cases, we aim at a low RMSE score, when compared to the previously implemented manual forecasting process.

Improving data-driven decision-making, as well as pricing strategies, are also part of the defined criteria for business success, as it is expected that the developed sales forecasting model can be able to deliver accurate and timely sales forecasts and insights, to better support both data-driven decision making and pricing recommendations, from an organizational point of view.

Finally, the improvement on identifying sales trends and patterns also englobes the defined business success criteria. By providing accurate identification of sales trends and patterns, such as seasonality, product demand and customer behaviour, Siemens will be able to incorporate them in their business strategy, which can ultimately translate into increased revenues and profit margins.

2.4. SITUATION ASSESSMENT

The situation assessment phase of the project encloses detailed analysis of resources, risk assessment, and cost-benefit analysis, that should be encompassed when preparing the project plan and related goals. This step also includes considerations on the provided data for analysis, as well as on the intrinsic business context in which the project is inserted.

The initially provided data for analysis consisted in two datasets: “Case3_Sales_data.csv” and “Case3_Market_data.csv”. The firstly indicated dataset was composed of 3 columns and 9802 rows, containing sales transactional data from 2018 to 2022, such as transaction date, amount, and product category. The secondly indicated dataset had 48 columns and 222 rows and contained data on diverse macroeconomic indices relevant for Siemens activity, correspondent to different countries and dates, from 2004 to 2022.

The data was subsequently explored and prepared, using for that purpose some of the most well-known Python libraries. For preparing and working details inside the notebook, *pandas* and *numpy* for preparing and were used, while both *matplotlib*, *seaborn*, and *colorama* were used for visualization matters. Regarding the trend and seasonality analysis only *statsmodels* was used. Lastly, for the data preprocessing both *scipy* and *xgboost* were used. The libraries chosen for the modelling step were the following: *sklearn, xgboost*, *statsmodels, pytorch-forecasting, neuralprophet* and *neuralforecast*.

In accordance with the guidelines, and in order to improve model output results, it was decided to include an external data source regarding Germany's GDP from the Federal Reserve Economic Data (FRED) database.

Developing such project also involves acknowledging possible barriers or events that might occur with impact on schedule, cost, or results, as well as identifying measures that can prevent them. This involves an effective risk and contingency plan assessment, improving the capacity and preparedness to mitigate such situations.

Poor data quality represents a potential risk, such as missing values, data inconsistencies or lack of data. For this purpose, diverse data cleaning techniques were carefully implemented to address this problem, in order to obtain a clean and complete pre-processed dataset ready to be used in the modelling step. Also in this context, ensuring a clean and complete pre-processed dataset avoids risk of major problems in model implementation.

Additionally, shortage of domain knowledge could lead to present inaccurate forecasting, as well as out-of-context or missing assumptions and important market trends. Therefore, extensive research and external data source retrieval was held to tackle such risk, helping to build domain knowledge that could assist and guide the developed work in the subsequent phases of the project.

Data diversity and representativeness problems may also arise as risk, in the sense that the provided sales dataset only concerns to sales data from 2018 to 2022, which may not be totally representative of complete sales history. As so, ensuring data richness and diversity, possibly requesting access to more data in the future, can help to overcome such risk.

Moreover, inaccurate communication of project findings and recommendations to Siemens' stakeholders may pose barriers to its implementation. Therefore, it is important to ensure that the project presentation is effective and business-oriented, clearly, and concisely addressing actionable findings and recommendations to the relevant audience.

Regarding the cost-benefit analysis of this project, there are potential direct and indirect costs associated to computing resources, data processing and modelling, training for interpretation and use of sales forecasting outputs, as well as to changes in strategic paradigm. With regards to benefits, this project could help to enhance sales forecasting accuracy, increase sales margins, improve overall operational efficiency, enhance pricing strategies, as well as transparency, and ultimately lead to a better global understanding of specific market trends.

2.5. Determine Data Mining goals

Defining data mining goals is a fundamental step for the development of this project, as it entails the technical outputs that enable the achievement of the previously defined business objectives.

One of the defined data mining goals was to minimize the Root Mean Squared Error (RMSE), which translates into precise and reliable forecasts from the chosen models. RMSE quantifies the Euclidean distance between forecasted predictions and the corresponding actual values, being used as one of the most prevalent and widely accepted error metrics. Despite the focus on RMSE performance metric for the above reasons, involving other performance evaluation measures in the analysis, such as MAPE

or MSE, ensures a thorough assessment of sales forecast model performance, and is also one of the considered data mining goals for this project.

The next data mining goal defined relates to uncovering the most relevant variables within the diverse product groups. In this context, the evaluation of macroeconomic indices' influence, as well as of other external events that can affect the existing product groups, is of crucial importance to such analysis, contributing to a better perception of each product group's response to external conjectural factors.

Furthermore, separate forecasting processes for each product group are also considered as one of the selected data mining goals, as it will be able to incorporate the most relevant machine learning techniques suitable to each case, according to the unique sales trends, behaviour, and specificities of each product group. Such approach has the potential to enhance sales forecasting processes, by recognizing the uniqueness and richness of characteristics of each product group. In this context, another of the selected data mining goals consists in exploring diverse forecasting model solutions for each product group, including time-series models and other Machine Learning algorithms, which can help to, once again, capture the uniqueness of their characteristics and determine the most effective model for each case.

Ensuring that the models developed can generalise and adapt correctly to new data is another important data mining objective for this sales forecasting project. Avoiding an excessively close fit to training model is essential for guaranteeing an adequately high predictive power and, therefore, better, and accurate model forecasting performance.

Moreover, performing parameter fine-tuning can also be encapsulated as one of the data mining goals for this project, since finding the optimal parameters in each product group modelling step can conduct to enhanced model performance. The optimal solution will be assessed by testing several combinations of parameters within the model, as well as by evaluating their respective performance, with the aim of reaching a satisfactory result that can ensure stable and high forecasting performance.

Lastly, developing a predictive solution that is both robust and scalable, as well as able to adapt to the constantly evolving nature of the data and high data throughput, is another data mining objective for this project. Such aspects will be put to consideration when choosing the appropriate model for each product group, in order to enable benefits that can include cost reduction and time-effectiveness.

3. METHODOLOGY

The data used throughout the project was made available by a couple of employees of the Siemens Advanta company making them real-time data and subjective to have some limitations.

3.1. DATA UNDERSTANDING

Siemens team, initially, made available three datasets which are 'Case2_Sales data.csv' (the sales data), 'Case2_Test Set Template.csv' (the test template) and 'Case2_Market data.xlsx' (the market data).

The sales dataset consists of 3 columns and 9802 rows with daily sales per product group (14 unique product groups). The data available covered the period of October 2018 to April 2022. Several problems were discovered in this dataset. Firstly, while renaming columns to make column names more meaningful and shorter, it was discovered that the 'Full_Date' column name holds meaningless symbols. Additionally, the 'Full_Date' column had the wrong data type format for the information it explores. Subsequently, the 'Sales €' column had the wrong separator for the decimal part which was not allowed to convert this column type into a float.

Moreover, Total sales by product groups show three main groups of products which are #1, #3 and #5 with a predominance of group #1 (Figure 1). Relatively to sales by the day, no big insight was possible, but the fact that some problems present some negative values, this possible due to the returns being higher than the sales of that same day. Grouping sales per weekday shows, that average sales are higher on Mondays and on Fridays (Figure 2). What is more, when grouping sales by product group, it shows that average sales of product group #1 are higher on Mondays (Figure 3), whereas average sales of product groups #3 (Figure 4), #5 (Figure 5) and some others are higher on Fridays (Figure 6). The average sales per month show that overall, the month that sells the most is September and the least is January, regarding specific products some like #1, #3, #4, and #5 follow a similar distribution and are relatively constant (Figures 7 & 8). Other products like #8, #11, and #12 show a big decrease in October after high sales in September (Figure 9). Product #13 does not sell very much in May (Figure 10). Product #14 shows an interesting pattern, where it increases within 3 months, but then the next month the sales are minimal and repeat the same pattern (Figure 11). Product #16 sells the most in December (Figure 12). Product #20 sells very well in April and in the last trimester (Figure 13). Product #36 had the highest sales in April and June (Figure 14).

Regarding the autocorrelation plots (Figure 15), it shows that the product has a few lags that are significant to explain the current sales, for example, product #5, where the lags 3, 6, and 10 are the most significant since they correlate the most with the current year sales. A test for stationarity was also made using the Augmented Dickey-Fuller Test, which showed that almost all product sales are stationary trends, except for product #8 which is not stationary and shows a crescent trend (Figure 16).

The market dataset has 48 columns and 222 rows. The first three rows are the header for the dataset so the first two rows were combined into one row while the third row was dropped. Regarding the 'Month Year' column, it was converted into the date-time format. Additionally, all columns with decimal values were cast to type float. After the described transformations were performed the dataset ended with 219 rows. Each row of data represents monthly indices that cover the period of February 2004 to April 2022.

Regarding the evolution of each macroeconomic indicator, many exhibit considerable variation between years, often demonstrating a discernible upward trend. Conversely, some indicators showcase a contrasting pattern with a declining trend over time. Notably, a significant downturn is observed across most macroeconomic features towards the end of 2008 through 2009, followed by a gradual recovery in 2010. This phenomenon is attributable to the global financial crisis that rocked economies worldwide, exerting profound impacts on macroeconomic dynamics. Similarly, a comparable pattern emerged at the onset of 2020, characterized by a sudden decrease in most macroeconomic indicators. This abrupt downturn is highly correlated with the onset of the COVID-19

pandemic, which unleashed unprecedented disruptions across global economies, thereby exerting significant downward pressure on macroeconomic metrics (Figure 17).

The test template consists of 3 columns and 140 rows that will be used for model evaluation. Some challenges were encountered while preparing the test template, particularly regarding the 'Month Year' column. It was found to contain extraneous symbols, which were subsequently removed. Additionally, the column included German month names, necessitating their replacement with English equivalents. Furthermore, there was a need to convert the column values into date-time objects to facilitate data processing.

3.2. DATA PREPARATION

The primary objective regarding the data is to ensure its readiness for the modelling phase. To accomplish this, the sales dataset undergoes segmentation into smaller datasets, each corresponding to specific product groups, supplemented by pertinent macroeconomic features.

To achieve the desired formatting, both datasets undergo a refinement process where their data types are optimized. For instance, converting float64 to float16 in the macroeconomic data. This optimization not only enhances efficiency but also facilitates time series forecasting with datetime variables.

Although after analysing sales data it did not reveal missing values, Market data revealed 53 rows with missing values. Conducting further analysis has been decided that data related to the period before the year 2016 would not be relevant for the project and was not exposed to missing values treatment, since the lag features only go as far as 1 or 2 years. So an analysis was made of every missing value after the mark of 2015, and a new pattern is shown, some features had missing values on the last month available only, others like 'UK Shipments Index M&E' and 'UK Producer Prices Electrical eq.' had missing values in the last 18 months, meaning that there was no data for the period of analysis from 2020/11 to 2022/04, these represent almost 1/4 of the whole period of analysis, it this in mind it did not make sense to use common tactics to treat missing values (like mean imputation, removal of the rows, or prediction according to other features), so the consideration was to remove these two features. All other variables with only one missing value, were filled with the mean of the last 3 available months (Figure 18).

When incorporating additional features, the impact of the COVID-19 lockdown on market dynamics, and subsequently on sales, was deemed significant. Consequently, it was decided to integrate this event into the sales data as a binary variable, denoted as 1 during periods following business lockdowns. Additionally, consideration was given to Brexit, recognizing its potential influence on trade negotiations between the UK and Germany, and thus its potential impact on products exchanged between the two countries. However, the proximity of the Brexit (contract finalization date) and the Covid-19 lockdown, occurring within a two-month window, resulted in a notably strong relationship between the two events, leading to a high correlation.

Another factor considered was Germany's GDP per quarter, intended to serve as a macroeconomic feature due to its potential impact on certain product sales. Unfortunately, this variable had to be excluded due to perfect collinearity with other variables. Speaking of collinearity, in this stage, any

feature exhibiting a correlation exceeding 95% with another feature was removed, resulting in the elimination of 15 features.

In the subsequent step, outlier treatment was applied to the sales data utilizing two distinct techniques: the Z-score method for sales exhibiting a normal distribution, and the Interquartile Range (IQR) or Box-plot method. For the Z-score method, the identification of products with sales following a normal distribution was imperative. Among these, only products #3, #5, and #6 were found to adhere to a normal distribution. Notably, only product #3 presented a missing value in January 2021 (Figure 19). Meanwhile, the IQR method involved calculating the interquartile range to assess upper and lower bounds, enabling the identification of outliers. A list of sales surpassing these limits was generated. Moreover, box plots were generated for each product group to visualize the distribution of sales (Figure 20). This visualization revealed an extremely low value for product #1, a heavily right-skewed tail for product #16, and two outliers each for products #3 and #5. However, given that these outliers conformed to a normal distribution, the Z-score criteria were prioritized. Given this context, despite the abundance of outliers identified through the IQR method, only specific actions were taken. The outlier represented by the lowest extreme value of product #1 and the outlier detected via the Z-score method for product #3 were addressed. For product #1, the outlier value was replaced with the mean of the sales data from the preceding 6 months. Regarding product #3, a threshold was established using the mean and standard deviation, with any values falling below three standard deviations considered outliers and subsequently adjusted.

After undergoing various transformations and treatments, the original dataset columns have been refined into the following features:

Table 3.2.1 – Final Features Table

Dataset	Feature		
Both	• Full_Date (Month Year)		
Sales	• CGK (product group)	• Sales €	• COVID_Impact
Market	• China Production Index M&E • France Production Index M&E • France Shipments Index M&E • Germany Production Index M&E • Germany Shipments Index M&E • Italy Production Index M&E • Italy Shipments Index M&E • Japan Production Index M&E	• US Production Index M&E • US Shipments Index M&E • Europe Production Index M&E • Europe Shipments Index M&E • Price of Base Metals • Price of Energy • Price of Natural Gas index • United States: EUR in LCU • US Producer Prices Electrical eq.	• Global Production Index Machinery eq. • Switzerland Production Index Machinery eq. • US Production Index Electrical eq. • Global Production Index Electrical eq. • Switzerland Production Index Electrical eq. • UK Production Index Electrical eq. • Italy Production Index Electrical eq. • Japan Production Index Electrical eq.

	• Switzerland Production Index M&E • UK Production Index M&	• France Producer Prices Electrical eq. • China Producer Prices Electrical eq.	• France Production Index Electrical eq. • Germany Production Index Electrical eq.

In the final stages of the analysis, the focus was on identifying the most significant lag for each macroeconomic index relative to the sales data of each product group. Simultaneously, the most significant sales lag for each product was pinpointed. Given the relatively low number of rows in each product sales dataset and to mitigate data loss, the lag analysis was limited to a maximum of 6 lags. Going deeper would necessitate the removal of more than 6 rows of data, which was deemed impractical. Although some product sales exhibited strong autocorrelation with lags around the 10-month mark, adding such lags would result in losing 10 or more rows of data, which was considered unreasonable. Recognizing the importance of feature selection for optimal model performance, two methods were employed to identify good features for each product sales dataset. Firstly, Spearman correlation was utilized to select features with a correlation higher than 40% with sales, while also ensuring they were not highly correlated with any other feature. Additionally, the importance plot generated by XGBoost was used to select the top 10 features deemed most influential in explaining variations in sales. These rigorous feature selection techniques ensured the inclusion of the most pertinent variables for accurate model predictions.

3.3. MODELLING

In this step, various modelling techniques are applied to the time series dataset to build forecasting models. Traditional machine learning algorithms such as linear regression, XGBoost, random forest, and decision trees were employed alongside advanced time series forecasting models. For traditional machine learning algorithms, k-fold, cross-validation, and grid search were used to optimize hyperparameters and assess model performance.

However, it is important to note that despite the efforts, these machine learning models faced challenges in accurately forecasting the time series data. This was primarily due to the lack of future macroeconomic features, which are crucial for making reliable predictions. Without access to comprehensive and timely macroeconomic data, the machine learning models were unable to effectively capture the complexities of the time series data.

Furthermore, it became apparent that the success of forecasting models heavily relied on the availability of future macroeconomic features, which were not readily accessible or predictable within the scope of this analysis. As a result, the hypothesis regarding the effectiveness of machine learning models for time series forecasting was challenged, and the feasibility of leveraging these models within the given constraints of time and data availability was called into question.

In addition to exploring traditional machine learning approaches, specialized time series forecasting models such as Neural Prophet, ARIMA (AutoRegressive Integrated Moving Average), and SARIMA (Seasonal AutoRegressive Integrated Moving Average) were also evaluated. These models are specifically designed to capture the temporal dependencies and patterns inherent in time series data.

3.4. EVALUATION

The evaluation step focused on assessing the performance of the developed forecasting models using the Root Mean Square Error (RMSE) metric, as preferred and required by the client. RMSE was chosen as it provides a measure of the average magnitude of the errors between predicted and observed values, thus indicating the accuracy of the forecasts.

Each forecasting model, including both traditional machine learning algorithms and specialized time series models, was evaluated using RMSE. This metric allowed for a direct comparison of the models' performance in terms of predictive accuracy. Models with lower RMSE values were considered to have better predictive performance, indicating a closer alignment between predicted and actual values. Therefore, in the following table it is specified the models that should be used by product group.

Table 3.4.1 – Final Scores per Group Product

Product	Model	Score	Product	Model	Score
Product #1	Neural Prophet	2260577.91	Product #11	SARIMA	141456.3
Product #3	Neural Prophet	1480871.38	Product #12	Neural Prophet	121590.02
Product #4	ARIMA	122955.48	Product #13	Neural Prophet	13585.74
Product #5	Neural Prophet	3206408.8	Product #14	ARIMA	12761.56
Product #6	ARIMA	261838.5	Product #16	SARIMA	187291.23
Product #8	Neural Prophet	271645.88	Product #20	ARIMA	2326.14
Product #9	ARIMA	4740.82	Product #36	ARIMA	23323.46

Some notes regarding the scores results, is that some product groups were not optimized to their model, even though these were the models and results chosen. Additionally, some models presented better results in some cases (in terms of overall RMSE) but the line prediction was unrealistic or very incredible. For example, for in product #4 SARIMA presented better results than ARIMA but the difference between train and validations, where too much, so it would be even more difficult to be sure of how good the test results were. In another example, some products show a prediction for the next 10 months that is just a repetition of the previous months, or in another cases, most months were predicted to return negative values, and even though it was considered a possibility, the likelihood of it happening in a sequence of months, was very little.

4. RESULTS EVALUATION

Although each model was evaluated to have the best performance possible when used with the data, it is not possible to clearly state its performance without being applied in a real environment. Nevertheless, it is conceivable to have an overall quality of the model since it was evaluated using the company's actual historical data. For example, it was created based on the guidelines that the customer established which are believed to be already studied and consistent with the results that the company expects. Moreover, when the company applies these models to their latest data, they should have specialised people from different departments, namely, marketing, customer, IT, and sales. This way they can evaluate the quality of the model and tailor-make them to fit their purposes and to retain

as much information out of it. What is more, if these work for the German division, this solution could be applied to other countries and make a performance comparison cross country.

Subsequently, it is believed if the models test on real-time data has a good performance, then it means that all business objectives were accomplished. Consequently, the company has a lot of insights that will help in the process of decision-making, improve resource allocation which leads to lower prices that is a key factor into having competitive advantage. Furthermore, now that the company has data transparency, it is easier to understand customers and so find new ways to keep and capture clients.

5. DEPLOYMENT AND MAINTENANCE PLANS

Coming from a complete development of the sales forecasting solution, it is important to ensure that an effective planning for the deployment into production, as well as a maintenance plan that can remain accurate and relevant over time, are properly implemented, to deliver value to the company. In this context, it is essential to work closely with all involved stakeholders to assess the needed requirements for model deployment into production, and the processes required for monitoring and maintaining the model on a time basis.

There were identified 4 relevant steps for the deployment and maintenance plans of the developed model. The first step involves presenting the developed model to the key staff involved, to ensure alignment with the business demands and needs. Depending on feedback received, it may be necessary to return to the model development phase until the company's expectations are fulfilled, and practical results can be delivered.

The second step involves model deployment into the existing sales and operations planning system (S&OP), integrating it into the existing workflows and technological processes, and performing compatibility testing within Siemens' digital infrastructure, as well as the relevant adjustments in the event of system conflicts.

Furthermore, the third step refers to support and training for involved staff, in order to enable them to effectively interpret model outputs, as well as to use the forecast model results at service of production, inventory, and sales strategies' decision-making. Additionally, guidance will be provided to keep up-to-date employee's capacities to manage and update the sales forecast model over time.

The fourth step both concerns continuous monitoring and maintenance of the developed model to assess and enhance performance, including re-training and regular updating with new data associated to changes in sales trends and patterns, as well as adjusting parameters and re-monitoring performance based on the comparison between predictions and actual sale results. In this way, we recognize the need for model accountability for market changes and variations in the economic landscape over time, pointing out potential improvement possibilities that can be turned into necessary adjustments. Regular documentation on model's performance will also be produced to track its behaviour over time regarding to data updates, as well as to record the relevant findings.

Monitoring will be conducted to ensure that the sales forecast model remains accurate and relevant. This will be achieved by tracking specified key performance metrics, such as Root Mean Square Error (RMSE), that will be compared to the previously defined business success criteria. Also in this spectrum, the cost-benefit relationship of the deployment and maintenance plan will also be continuously

assessed, by monitoring costs related to acquisition and preparation of new data and model updates, while also assessing the outcome benefits resulting from the successful implementation of the model, such as improved sales, inventory management, and reduced costs, for the sake of a more efficient production planning process. These practices will globally contribute to the optimization of deployment and maintenance plan, whilst ensuring the economic sustainability and viability of the sales forecast model.

6. CONCLUSIONS

To summarise, the adoption of the previously mentioned recommendations shall lead to business success. Improving the model's accuracy provides more granularity to lead the company to understand sales patterns which allow to anticipate market demand and so refining decision-making. Moreover, sales forecasting improves resource allocation which provides a wider profit margin which will allow the company to practise lower prices that are assumed to make the company to gain market competitive advantage. Furthermore, to develop the forecasting model, it was of the utmost importance to understand the data's transparency that makes the information acquired more trustworthy. Additionally, after having the right tools, Siemens has the tools to better understand its clients and define customer segmentations that will allow personalisation.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Nowadays, markets are constantly changing so it is crucial that Siemens does it as well. Consequently, it is recommended that the company keeps on investing in more robust systems in order to collect as much data as possible since, as mentioned before, the sales forecasting of the company are highly affected by macroeconomic factors and currently this information it is still too limited. Moreover, this should be aligned with having a data science team that clearly defines the required data and its characteristics, namely, features.

Although investing in cutting-edge technology and having a specialised team is essential to keep up with the market trends, it is of utmost importance that the company promotes cross-department collaboration in order to have multiple views on what is happening inside and outside the company. Let's take marketing campaigns as an example. If there is good communication between the data science and marketing departments, advertisement shall be more tailor-made which increases the percentage of customer retention.

Lastly, if the company keeps on updating its KPIs and follows the previously mentioned recommendations, it is believed that Siemens product sales will increase.

In conclusion, this project provides the tools to update Siemens sales product process in theory. However, it would be of Group's E interest to provide support in the next steps of applying this project in the company itself so how can we help Siemens to see this initiative approved and operational in their daily operations?

7. REFERENCES

TURNER, J. (2023, September 4). *Why did the global financial crisis of 2007-09 happen?* Economics Observatory. https://www.economicsobservatory.com/why-did-the-global-financial-crisis-of-2007-09-happen

Eurostat. (1991, January 1). *Gross Domestic Product for Germany*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/CPMNACNSAB1GQDE

How to disable Python warnings? (n.d.). Stack Overflow. https://stackoverflow.com/questions/14463277/how-to-disable-python-warnings

Pandas DataFrame: Show All Columns/Rows | Built In. (n.d.). Builtin.com. https://builtin.com/data-science/pandas-show-all-columns

COVID-19 in Germany: Back to Business after the Lockdown – Guidance for Employers | Insights | Mayer Brown. (n.d.). Www.mayerbrown.com. https://www.mayerbrown.com/en/insights/publications/2020/05/ger-covid19-in-germany-back-to-business-after-the-lockdown-guidance-for-employers

NeuralProphet documentation. (n.d.). Neuralprophet.com. https://neuralprophet.com/contents.html

Triebe, O. (n.d.). *neuralprophet: NeuralProphet is an easy to learn framework for interpretable time series forecasting.* PyPI. https://pypi.org/project/neuralprophet/

Forecasting. (n.d.). Open Time Series. https://opentimeseries.com/python_packages/forecasting/

8. APPENDIX

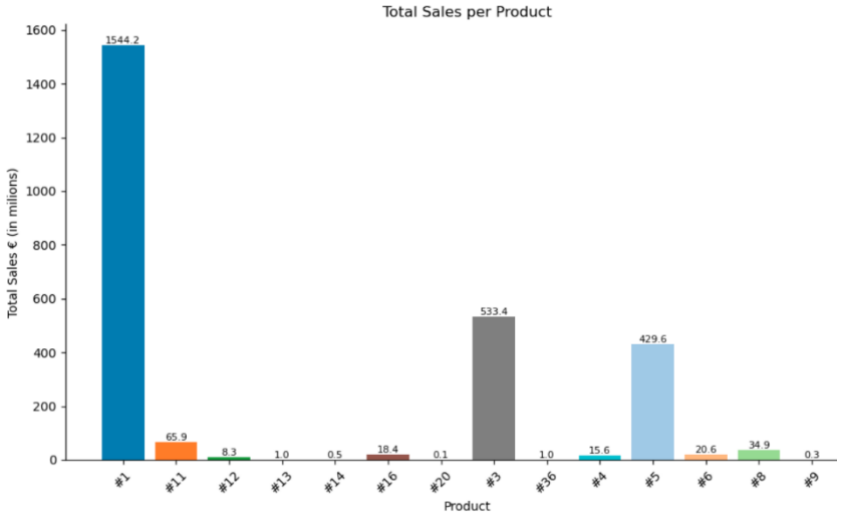


Figure 1 – Total Sales per Product

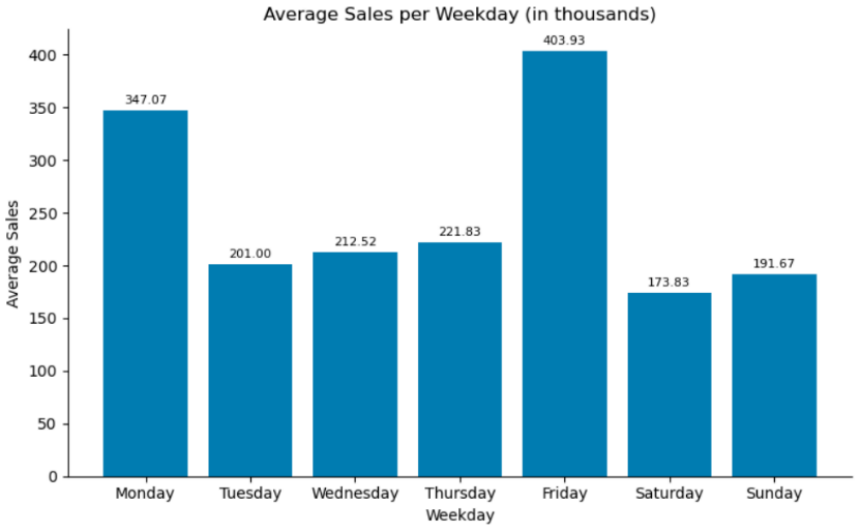
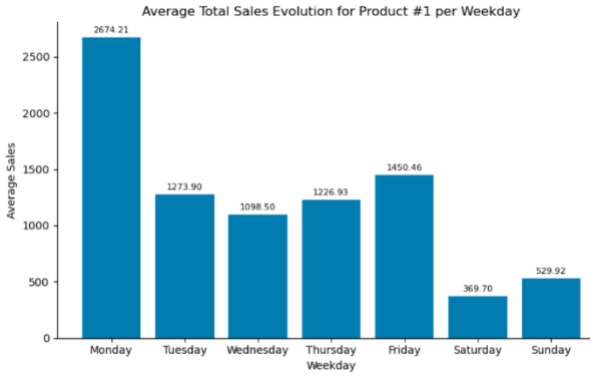


Figure 2 – Average Sales per Weekday



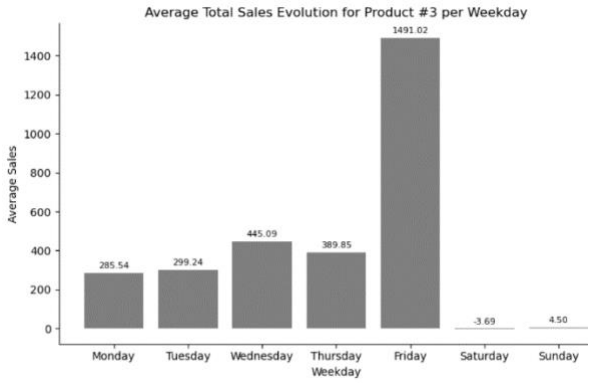
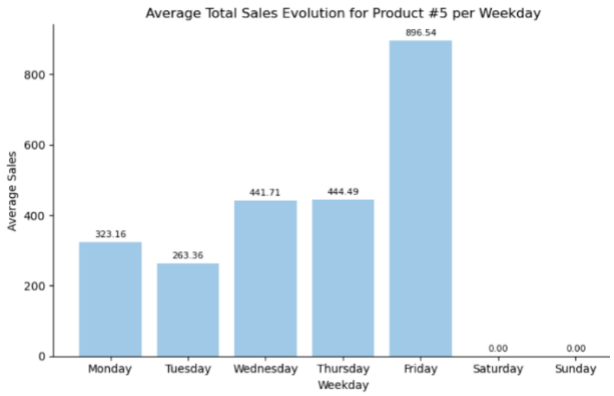
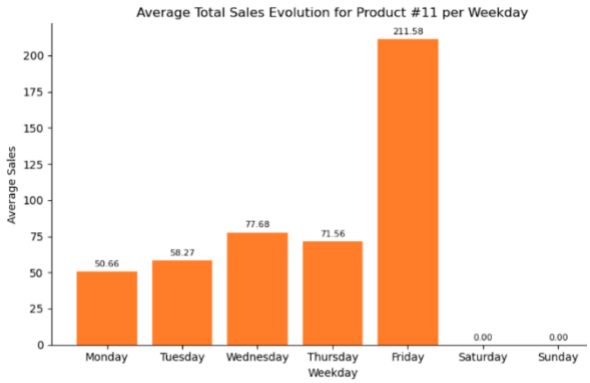


Figure 3 & 4 – Average Sales per Weekday and Product Group #1 and Group #3





Figures 5 & 6 – Average Sales per Weekday and Product Group #5 & Group #6

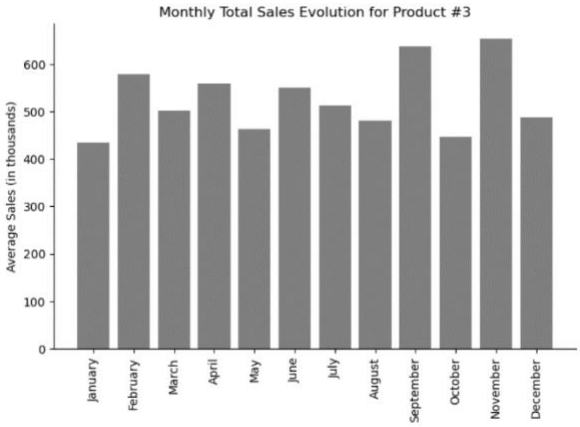


Figure 7 & 8 – Average Sales per Month and Product Group #1 & Group #3

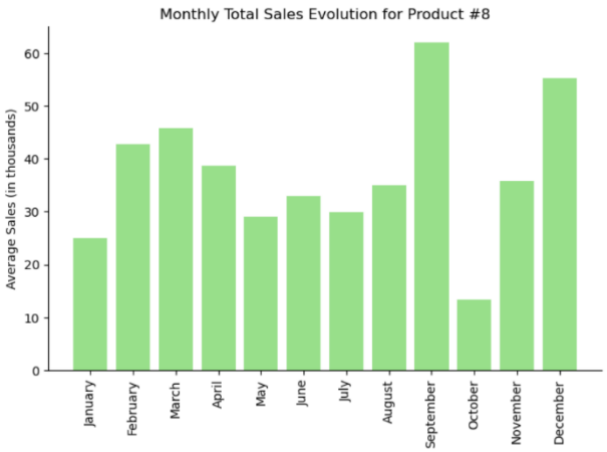
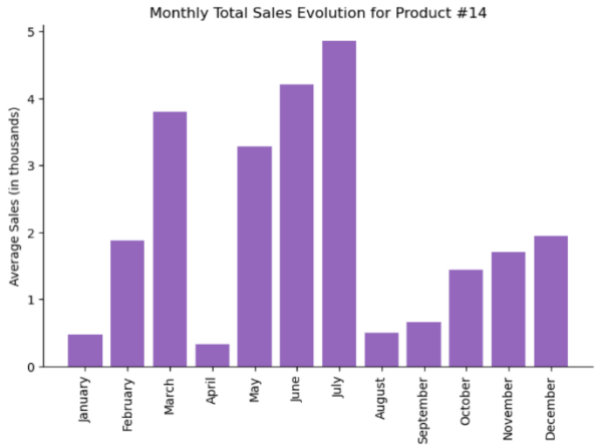
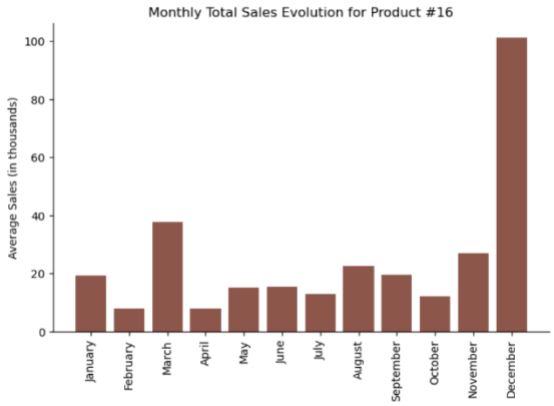
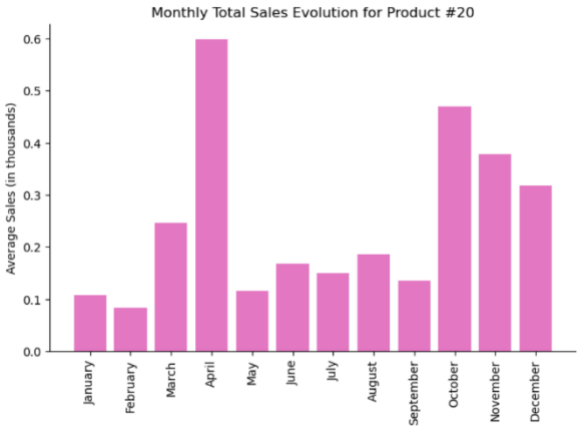


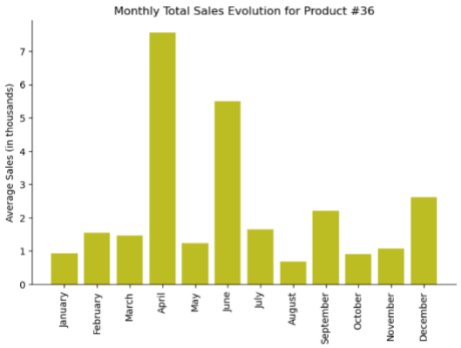
Figure 9 & 10 – Average Sales per Month and Product Group #8 & Group #13





Figures 11 & 12 – Average Sales per Month and Product Group #14 & Group #16





Figures 13 & 14 – Average Sales per Month and Product Group #20 & Group #36

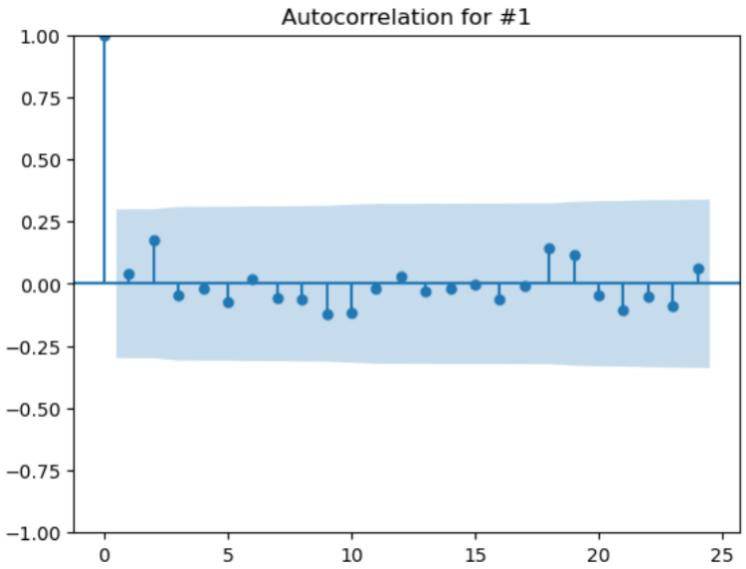


Figure 15 – Example of Autocorrelation plot

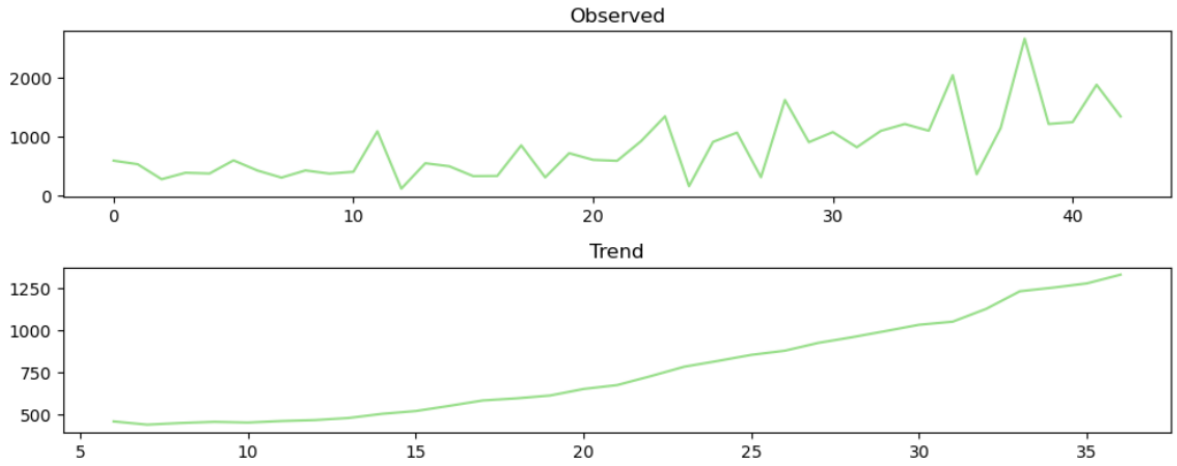


Figure 16 – Example of Evolution of Sales of Product #8 and its trend

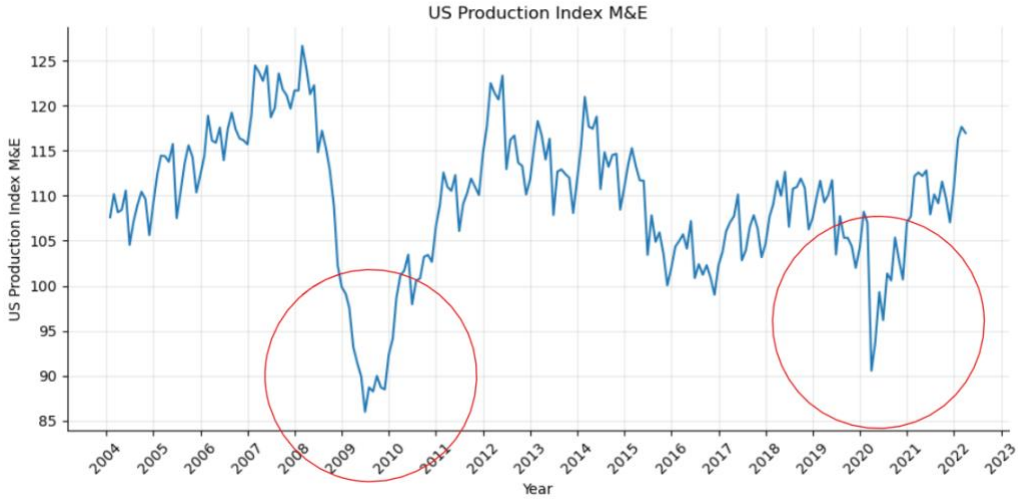


Figure 17 – Example of Macroeconomic evolution (case for US Production Index)

	Column with Missing	Number of Missing	Percentage
14	UK Shipments Index M&E	18	23.684211
27	UK Producer Prices Electrical eq.	18	23.684211
11	Switzerland Production Index M&E	1	1.315789
12	Switzerland Shipments Index M&E	1	1.315789
16	US Shipments Index M&E	1	1.315789
34	Switzerland Production Index Machinery eq.	1	1.315789
42	Switzerland Production Index Electrical eq.	1	1.315789

Figure 18 – Market data missing values for 2016-01 to 2022-04

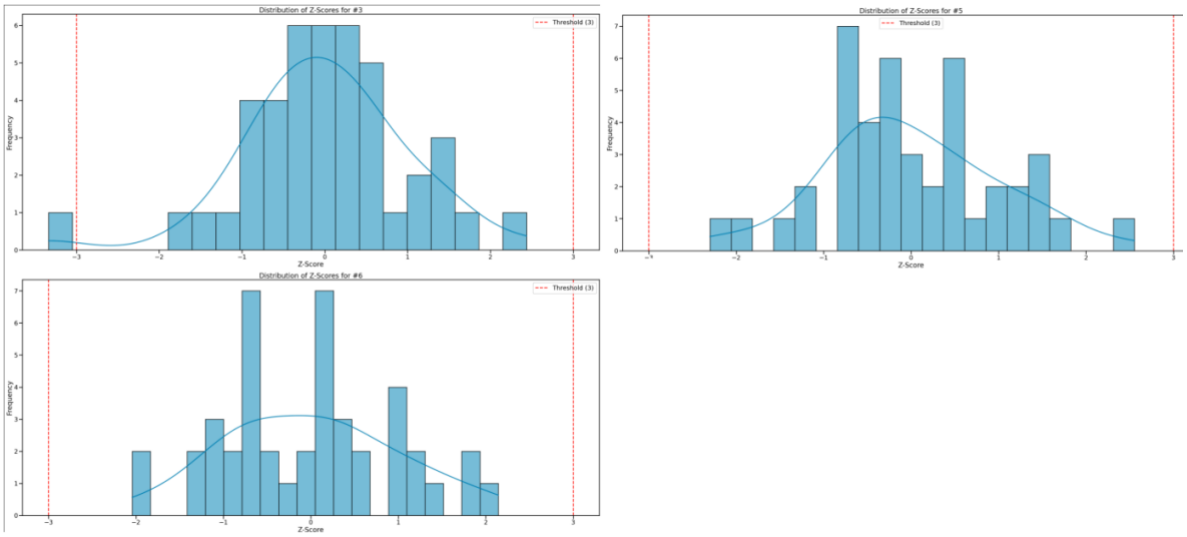


Figure 19 – Normal Distribution Plot of Normal Distributed Sales

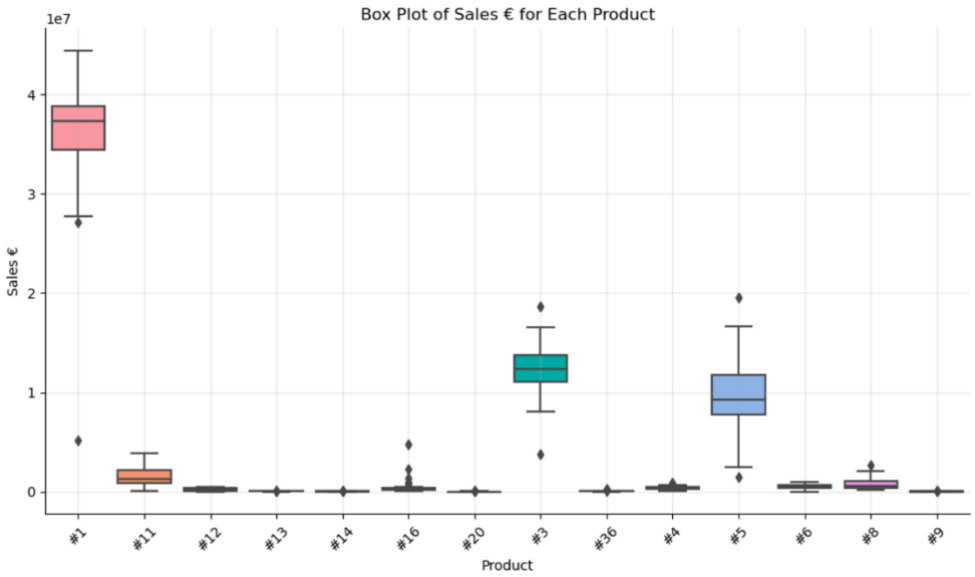


Figure 20 – Boxplot of Sales per each Product Group