Data-Driven Demand Forecasting for Pharmacy Supply Chains

Author: Apostolos Syrris (2739810)

Company supervisor: Tsaknakis George

First VU supervisor: dr. Rene Bekker

Second VU supervisor: dr. Sandjai Bhulai

A thesis submitted in fulfillment of the requirements for the MSc in Business Analytics (Financial Risk Management)

Abstract

The current thesis develops and evaluates a segment-aware, short-term demand forecasting framework for a para-pharmaceutical supply chain that is operating in the Greek pharmacy market. Regarding the forecasting problem it is characterised by a heterogeneous portfolio of 448 weekly SKU-level series spanning the period from 2022 to 2025, where demand ranges from stable and seasonal to sparse and intermittent, with more than 59% of all observations recording zero sales per week. Also the framework combines ABC–XYZ segmentation for model routing with five baseline forecasting methods: Seasonal Naïve, ETS (Error, Trend, Seasonality) , Croston, SBA(Sensitivities-based approach) and SARIMA, this model selection is driven by empirical WAPE minimisation at segment level, resulting in SARIMA being assigned to the stable AY segment and SBA to the intermittent AZ, BY, BZ, and CY segments where promotional and holiday effects are incorporated through two complementary mechanisms: exogenous regressors within SARIMA and post-hoc multiplicative uplift adjustments for SBA and Croston. The reason for the exclusion of CZ segments relies mostly on the limited data and many zeros which made it very difficult to make a forecast. This means that the final pipeline is evaluated on a fixed 12-week holdout horizon across 172 SKUs and 2,064 forecast observations. Results show that the SBA-routed segments achieve an average WAPE of 7.84% and MAE of 10.45 with external variables such as promos and holiday effects as mentioned above, while the SARIMA-routed AY segment achieves a WAPE of 17.02% and MAE of 28.63. The findings confirm based on the results they got that forecast accuracy is primarily driven by the underlying demand structure of each segment rather than by model complexity alone, and that a modular, segment-based routing strategy outperforms a one-size-fits-all approach for heterogeneous retail portfolios.

Contents

Chapter 1

Introduction

1.1 Background

The Greek pharmacy market is characterized by a large number of independent pharmacies and a distribution process that is still influenced more by practical experience than by formal forecasting systems, having that in mind for many non-prescription and para-pharmaceutical items, product availability is often managed through frequent replenishment and order-on-demand practices rather than systematic stock planning. This makes short-term demand forecasting relevant for both suppliers and pharmacies, as it directly affects inventory levels, replenishment timing, and service performance.

This particular study focuses on para-pharmaceutical SKUs (Stock Keeping Units) because they are operationally important and comparatively more accessible from a data perspective than prescription pharmaceuticals, which are subject to stronger regulatory and market constraints. At the same time, the demand for stock keeping products is not fully stable, as it may vary with seasonality, holiday periods, payday timing, and promotional activity. This means that these characteristics make the category suitable for examining how short-term forecasting methods perform under different demand conditions.

The empirical analysis is based on a multi-series weekly dataset covering several years and enriched with promotion-related and calendar-based information meaning that the practical objective is to support and give a better understanding for the sales and decisions in order to reduce operational inefficiencies, while the scientific objective is to evaluate forecasting performance across product groups with different demand patterns.

1.2 Problem Statement

In the Greek pharmacy channel, replenishment decisions for para-pharmaceutical products are often made under uncertainty and with limited operational flexibility. At store level, ordering still depends heavily on local judgement about what is likely to sell rather than on systematic demand forecasts, while sometimes sales depend on the number of visits of each independent seller. As a result, product availability is uneven: when customers request specific items, pharmacies may not always have them in stock and may instead offer substitutes or place delayed orders which can lead to customer inconvenience and, in some cases, lost sales.

The problem is amplified by the operational characteristics of para-pharmaceutical SKUs. Pharmacies typically face limited storage capacity expiration constraints and relatively high holding risk for products that may be expensive or difficult to return meaning that conservative ordering can reduce availability and sales, while aggressive ordering can increase excess inventory, tied-up capital, and expiry-related losses. At the distribution level, unstable ordering patterns also create inefficiencies since urgent replenishment orders increase delivery complexity and cost while low order density may reduce transport efficiency.

From a forecasting perspective the challenge is that the portfolio contains many SKUs with different demand regimes including sparse and intermittent series with frequent zero-sales weeks. Demand is also influenced by external drivers such as promotions and holidays. These effects may cause temporary deviations from the underlying demand pattern which simple seasonal models do not capture well while the central analytical challenge is therefore to develop a forecasting approach that can handle heterogeneous SKU behaviour incorporate relevant external effects and avoid imposing a single model structure on the entire portfolio.

A suitable solution is a standardized forecasting and evaluation procedure that produces consistent short-term SKU-level estimates and supports operational planning under these constraints. Effectiveness is assessed through backtesting against a simple benchmark, with performance reported by SKU segment and forecast horizon, so that the results remain interpretable and operationally relevant.

1.3 Research Questions

This study focuses on one main research question below and several sub-questions that break down the key operational and methodological issues into specific points that can be tested.

Primary research question (RQ1):

- RQ1: How can Just Health design and evaluate a short-term demand forecasting framework for its para-pharmaceutical portfolio, and how does forecast accuracy vary across product groups with different demand patterns?

Sub-questions:

- RQ1.1: What is the forecasting accuracy of a simple benchmark (e.g., Seasonal Naïve) for the 8–12-week horizon, and which SKU groups are poorly served by this baseline?

- RQ1.2: To what extent do calendar-driven effects (seasonality and major holiday periods) and to what extent they contribute to forecast accuracy?

- RQ1.3: How much incremental predictive value is obtained by incorporating promotion information compared to using sales history and calendar effects alone?
- RQ1.4: How does forecast performance differ between SKU groups with stable seasonal demand versus intermittent/zero-inflated demand, and which modelling approaches are more robust for each group?
- RQ1.5: Does segmentation by sales importance and demand behaviour improve the overall planning usefulness of forecasts compared to applying a single method uniformly across the portfolio?

1.4 Thesis Outline

The remainder of this report is structured as follows. Chapter 2 reviews the relevant forecasting literature, with emphasis on short-term demand forecasting, intermittent demand, and the role of external drivers such as promotions and calendar effects. Chapter 3 describes the data used in this study, including data collection, preprocessing, and exploratory analysis through visualizations and descriptive statistics. Chapter 4 examines the available explanatory variables and discusses the feature selection process, with particular focus on promotions, holidays and calendar effects, lag-related effects, price-related variables, and segmentation. Chapter 5 presents the methodology, including the theoretical background of the benchmark and candidate models, the evaluation metrics, the train-test design, and the forecasting pipeline used in the implementation. Chapter 6 reports the baseline and final forecasting results, compares model performance across demand segments, and discusses the practical interpretation of the accuracy measures. Finally, Chapter 7 concludes with a discussion of the main findings, limitations of the study, and practical recommendations for future work.

Chapter 2

Literature Review

2.1 Demand Forecasting

Demand forecasting is a central task in supply chain and inventory planning because forecast quality directly affects replenishment decisions, service levels, and operating costs. In practical business settings, products often exhibit different demand behaviours, ranging from stable seasonal patterns to sparse and intermittent demand. Consequently forecasting research distinguishes between regular seasonal time series and intermittent demand series, since these patterns require different modelling assumptions and evaluation approaches (Hyndman and Athanasopoulos 2021; Fildes et al. 2022).

For regular demand series classical time-series methods such as exponential smoothing and ARIMA-type models remain widely used because they are interpretable computationally efficient and often competitive in applications. Furthermore exponential smoothing methods are suitable when demand contains level or trend or seasonal components, while ARIMA and SARIMA models are suited to series with autocorrelation and seasonal dependence (Hyndman, Koehler, et al. 2002; Hyndman and Khandakar 2008). Seasonal Naïve is also widely used as a benchmark for seasonal data, because it provides a transparent reference forecast by repeating the observation from the same season in the previous cycle. Benchmarking against such simple methods is needed because any proposed model should demonstrate added value relative to an understandable baseline rather than only being evaluated in isolation (Hyndman and Athanasopoulos 2021; Makridakis et al. 2022).

A separate stream of literature addresses intermittent demand forecasting where demand occurs sporadically and many periods have zero demand. Based on this sentence Croston's method is the classic approach, as it estimates demand size and demand interval separately before combining them into a per-period forecast (Croston 1972). However, Croston's original estimator is known to be biased which led to modifications such as the Syntetos–Boylan Approximation (SBA) and later variants designed to reduce bias and improve performance (Syntetos and Boylan 2005). Empirical studies show that the performance of these methods depends on the degree of intermittency and demand variability and also stability of the non-zero observations (Kaya et al. 2020).

Because demand series differ both in business importance and in forecastability, segmentation approaches are frequently used to support model choice and inventory policy design. ABC classification ranks items by value contribution when XYZ-type classification captures demand variability and stability. Having that in mind and combining these dimensions helps distinguish high-value stable items from lower-value or highly irregular items and supports differentiated forecasting and planning strategies (Flores and Whybark 1986). In a forecasting context segmentation allows the modeller to assign methods that better match the demand pattern instead of forcing one universal model on the entire SKU portfolio (Syntetos, Boylan, and Croston 2005).

Overall, the literature suggests that robust demand forecasting requires appropriate model selection, simple benchmarks, and evaluation across relevant demand segments. This is particularly important in SKU-level applications, where a portfolio may contain both smooth seasonal series and highly intermittent demand patterns. (Bergmeir et al. 2018; Hyndman and Koehler 2006).

2.2 Intermittent Demand

Intermittent demand refers to demand series in which sales occur irregularly over time with many periods of zero demand. This pattern is common in inventory and SKU-level circumstances for low-volume items that someone can find in pharmachy stores, spare parts, and products with unpredictable purchasing behaviour (Croston 1972; Kaya et al. 2020). This comes to contrast of regular seasonal series that intermittent demand is defined by two related elements: the size of demand when it occurs and the interval between non-zero observations. This makes forecasting more difficult with standard time-series models, which typically assume a more continuous demand process.

The main challenge in intermittent demand forecasting is that long stretches of zero observations reduce the usefulness of methods that rely heavily on recent demand levels or smooth seasonal patterns. Accordingly, classical exponential smoothing or standard ARIMA models may perform poorly when the demand process is sparse and irregular (Willemain et al. 1994). This has as effect that the modeller may need methods that can separately handle demand occurrence and demand magnitude, rather than treating the series as a conventional continuous process.

Croston-type methods estimate intermittent demand by separately modelling demand size and the interval between non-zero observations. SBA extends this approach by correcting the bias of the original estimator while maintaining its practical structure (Croston 1972; Syntetos and Boylan 2005).

Intermittent demand forecasting is not only a technical modelling problem but also a classification and method-selection problem. The more irregular the series becomes the more it is needed to use methods that are robust to zero-heavy patterns and can still provide stable forecasts for planning. Thus, intermittent demand series are often treated as a separate class in forecasting studies and are evaluated independently from smoother seasonal series (Syntetos, Boylan, and Croston 2005; Kaya et al. 2020).

Also, intermittent demand is particularly relevant in SKU-level forecasting because sparse or zero-heavy sales patterns require methods that differ from those used for regular demand series. Therefore, Croston-type approaches are considered alongside classical forecasting methods depending on SKU characteristics.

2.3 Promotions and External Demand Drivers

Demand forecasting is often influenced not only by the internal temporal structure of sales data but also by external factors that temporarily change purchasing behaviour. Among the external drivers are promotional activities, price reductions, holiday periods, and other calendar-related events. These variables can create short-term spikes or distortions in demand that are not explained by historical sales patterns alone. This has as a result models that ignore such effects may underpredict demand during promotional periods or fail to capture recurring event-related changes (Fildes et al. 2022; Abolghasemi et al. 2020).

Also, it needs to be mentioned that promotional effects are especially relevant in retail and consumer goods environments, where demand can respond strongly to campaigns, flyers, bundle offers, price reductions, or in-store visibility actions. Based on that, promotions may affect both the timing and the magnitude of purchases which lead to demand shifts that are temporary . The retail forecasting literature treats promotions as explanatory variables in causal or regression-based models, or as uplift adjustments applied to a baseline forecast when sufficient historical promotion information is available (Fildes et al. 2022; Bojer et al. 2019).

Holiday effects operate in a similar way, although their impact is usually driven by calendar structure rather than commercial intervention. Certain weeks may experience higher or lower demand due to public holidays, seasonal shopping behaviour or changes in store and logistics operations. These effects can be particularly relevant in weekly forecasting settings, where one holiday week may meaningfully alter the observed demand level. Including holiday indicators helps the model distinguish between normal demand variation and calendar-driven deviations (Huber and Stuckenschmidt 2020).

The inclusion of external demand drivers is notably useful when the forecasting problem involves heterogeneous products and operational planning at SKU level. The same product may follow a stable baseline pattern most of the time but still react strongly to promotions or holidays. Furthermore, combining historical demand information with external event variables can improve interpretability. In other terms these drivers can be incorporated either through exogenous regressors in statistical models, such as regression with ARIMA errors, or through post-forecast adjustment mechanisms that modify a baseline prediction according to expected event effects (Hyndman and Athanasopoulos 2021; Abolghasemi et al. 2020).

Below, promotions and holidays are treated as important demand drivers because they help explain deviations from the normal sales pattern and improve the realism of the forecasting framework. Their role is not to replace the core time-series structure of the model, but to complement it with event-specific information that is meaningful. This means promotion and holiday variables should be used where they add explanatory value, while simple univariate benchmarks remain necessary for comparison and for SKUs with insufficient external-event history.

2.4 Forecasting Evaluation Metrics

Forecast evaluation is a necessary part of demand forecasting because it determines whether a proposed method provides improvement over simpler alternatives. In forecasting the objective is not only to fit historical data but to estimate how well the model would perform when used for future planning decisions so evaluation should imitate the real forecasting situation as closely as possible and should be based on out-of-sample performance rather than only in-sample fit (Hyndman and Athanasopoulos 2021; Bergmeir et al. 2018).

This requirement is remarkably strong in SKU-level forecasting, where the portfolio may contain stable, seasonal, erratic, and intermittent demand series at the same time. A model that performs well on average may still fail for specific SKU groups, while a simple method may be sufficient for some segments but not for others. Therefore, evaluation should be reported both at the portfolio level and by relevant demand segments, so that accuracy results can be interpreted in business terms rather than only as overall statistical averages (Fildes et al. 2022; Hyndman and Koehler 2006).

A transparent benchmark is the starting point for meaningful evaluation. Seasonal Naïve is used for seasonal weekly data because it provides a simple forecast based on the same period in the previous seasonal cycle while forecasting competitions and applied forecasting studies show that simple benchmarks can be surprisingly competitive when many series are forecasted simultaneously (Makridakis et al. 2022). As a result more complex models should be accepted only if they improve on this baseline in a consistent and interpretable way.

Classical models such as ETS, SARIMA, and Croston-type methods should be evaluated against the same benchmark and under the same testing design. This ensures that differences in performance are due to the models themselves rather than to different data splits or inconsistent assumptions. It also supports fair comparison between methods designed for different demand patterns, such as ETS/SARIMA for more regular series and Croston/SBA for intermittent series (Hyndman and Athanasopoulos 2021; Syntetos and Boylan 2005).

Rolling-origin backtesting is widely recommended because it evaluates forecasts from multiple historical points rather than relying on a single train–test split. The model is repeatedly trained using available information and evaluated over future horizons, providing a more reliable estimate of forecasting performance. (Bergmeir et al. 2018; Hyndman and Athanasopoulos

2021).

The choice of accuracy metric also affects how results are interpreted. MAE expresses the average absolute forecast error in the same units as the data, which makes it easy to understand for users. On contrary , MAE is scale-dependent and cannot always be compared directly across SKUs with very different demand volumes. Percentage-type measures are often used to express error relative to actual demand, but some percentage metrics can behave poorly when demand is zero or close to zero (Hyndman and Koehler 2006).

For this reason, WAPE is often used in business forecasting because it compares the total absolute forecast error with the total realised demand over a set of observations. This makes it easier to report accuracy at portfolio or segment level and avoids some of the problems of calculating a separate percentage error for each zero-demand period. WAPE is also useful for communicating performance to stakeholders because it expresses forecast error relative to sales volume, while MAE remains valuable as a complementary unit-based measure (Kolassa and Schütz 2007; Hyndman and Koehler 2006).

In intermittent demand settings, evaluation is more difficult because zero-heavy series can produce unstable or misleading percentage errors. A model may appear accurate simply by predicting low values when demand is absent, while still failing to anticipate non-zero demand events. Because of this, intermittent-demand methods should be evaluated carefully and, where possible, separately from smoother demand series. Segment-level reporting helps reveal whether Croston-type methods actually improve forecasting for sparse SKUs or whether aggregation is more suitable (Syntetos and Boylan 2005; Kaya et al. 2020).

Therefore, this thesis evaluates forecasting approaches using rolling-origin backtesting, Seasonal Naïve benchmarks, and WAPE/MAE metrics. Results are analysed across ABC–XYZ segments and forecast horizons to identify which methods are most suitable for different SKU behaviours.

Chapter 3

Data

3.1 Data Collection

The analysis is based on the company's internal sales records of 448 SKUs covering the period from April of 2022 to the end of 2025. Annual transaction files were used as the primary source for reconstructing sales activity at SKU level, while the weekly sales extract provided the time-series structure required for the forecasting analysis. The annual files covered the full product universe through stable SKU codes and product descriptions, together with sales values and quantities, whereas the weekly dataset provided a standardized sequence of week-level observations for each item across the study horizon.

3.2 Data Description

The final dataset is organized at weekly level as a panel of SKU-week observations, for that, each record contains a unique SKU identifier, a week or date reference, product and segment information, the observed sales value used as the target variable, and additional promotional and calendar-based event indicators. This structure supports a consistent tracking of each product across the full study horizon from 2022 to 2025 as mentioned before.

The dataset is heterogeneous and includes both high-volume and intermittent-demand items, as well as zero-sales observations for some SKUs in specific periods. In addition to the weekly series, the underlying data sources provide complementary product-level and time-based information that helps preserve continuity across the portfolio and supports the later forecasting analysis.

3.3 Descriptive Statistics of the Dataset

The following table presents the descriptive statistics of the weekly SKU-level dataset used in this study. The dataset covers 448 SKUs observed at weekly frequency from March 2022 to December 2025, resulting in 48,245 observations across 7 ABC–XYZ demand segments. Statistics are reported at segment level and include demand scale, sparsity, and event coverage measures.

Table 3.1: Descriptive Statistics by Demand Segment

Segment	**SKUs**	**Obs.**	**Mean**	**Median**	**Std.**	**Max**	**Zero Ratio**
AY	28	4,836	84.28	49.28	102.24	995.60	0.099
AZ	62	9,555	54.60	15.00	125.16	1920.80	0.388
BY	2	185	25.16	15.00	52.20	550.00	0.173
BZ	82	11,785	14.18	0	30.44	617.00	0.506
CX	20	32	42.83	30.40	37.76	170.88	0
CY	3	20	72.39	63.25	42.56	141.70	0.100
CZ	251	21,832	2.50	0	10.88	268.80	0.846
Total	**448**	**48,245**	24.01	0	72.29	1920.80	0.594

Note: Mean, Median, Std., and Max refer to weekly sales in units. Zero Ratio is the proportion of weeks with zero recorded demand. Observations (Obs.) represent the total number of weekly records per segment. The dataset covers the period from March 2022 to December 2025.

3.4 Data Preprocessing

The preprocessing stage was designed to transform the raw sales files into a consistent weekly dataset suitable for descriptive analysis and forecasting while the main objective was to create a product-week panel in which each SKU could be followed over time using a stable identifier. Also, the key is to preserving important operational information such as zero-sales weeks, return-related entries, and calendar effects. The following section describes the cleaning and standardisation of the raw data, the weekly aggregation process, and the treatment of missing values and zero sales.

3.4.1 Cleaning and Standardisation

The first preprocessing step involved aligning the available sales records across years and standardising the structure of the dataset and since the raw files were organised monthly, the columns had to be checked and harmonised before the data could be combined into a single modelling dataset. The SKU code was used as the primary identifier, allowing the same product to be traced longitudinally even when naming conventions were not fully consistent across files.

A key part of this stage concerned operational entries that did not directly represent regular sell-through. In particular, the raw data contained zero-value records and negative weekly values associated with returns, corrections, or accounting adjustments. These observations were not automatically removed from the dataset, because they represent part of the company's commercial history. Instead, they were explicitly flagged during preparation so that non-standard movements could be identified and considered in later modelling decisions.

The final cleaned dataset preserved the original sales and introduced a more standardised structure for analysis. Each product was linked to its SKU code, the relevant time information was converted into a consistent date format, and special cases such as negative entries were kept visible through additional indicators in order to ensure that the preprocessing did not artificially simplify the data, while still making it suitable for forecasting and segmentation.

	A	B	C	D	E	F	G	H	I	J	
1	SKU	week_start_date	year	week_num	Περιγραφή	Προμηθευτής	Brand	Κατηγορία	Υποκατηγορία	Πωλήσεις	holiday_name
2	100	2022-05-16	2022	20	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	415	
3	100	2022-05-23	2022	21	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	264	
4	100	2022-05-30	2022	22	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	85	
5	100	2022-06-06	2022	23	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	36	
6	100	2022-06-13	2022	24	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	8	Holy Spirit Monda
7	100	2022-06-20	2022	25	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	21	
8	100	2022-06-27	2022	26	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
9	100	2022-07-04	2022	27	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
10	100	2022-07-11	2022	28	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
11	100	2022-07-18	2022	29	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	244	
12	100	2022-07-25	2022	30	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
13	100	2022-08-01	2022	31	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
14	100	2022-08-08	2022	32	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	17	
15	100	2022-08-15	2022	33	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	Dormition of the V
16	100	2022-08-22	2022	34	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	23	
17	100	2022-08-29	2022	35	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	39	
18	100	2022-09-05	2022	36	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
19	100	2022-09-12	2022	37	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
20	100	2022-09-19	2022	38	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
21	100	2022-09-26	2022	39	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	26	
22	100	2022-10-03	2022	40	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	23	
23	100	2022-10-10	2022	41	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	96	
24	100	2022-10-17	2022	42	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	50.76	
25	100	2022-10-24	2022	43	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	71.44	Ochi Day
26	100	2022-10-31	2022	44	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
27	100	2022-11-07	2022	45	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	252.67	
28	100	2022-11-14	2022	46	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	154.61	
29	100	2022-11-21	2022	47	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
30	100	2022-11-28	2022	48	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	
31	100	2022-12-05	2022	49	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	166.04	
32	100	2022-12-12	2022	50	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	154.16	
33	100	2022-12-19	2022	51	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	Christmas Day
34	100	2022-12-26	2022	52	SET THERASOL SOLUTION 250ml 1+1 ΔΩΡΟ		ABC	THERASOL	ΟΔΟΝΤΟΚΡΕΜΕΣ	0	New Years Day | S

Figure 3.1: Preview of the dataset

3.4.2 Weekly Aggregation

After the cleaning and standardisation stage, the sales data were aggregated to a weekly frequency. The weekly level was selected because the forecasting framework aims to support short-term operational planning and replenishment decisions. Each transaction or sales record was assigned to a week start date and week number, after which the data were grouped by SKU and week. This generated a continuous chronological series of weekly sales values for each product.

The final weekly modelling dataset represents sales history at the product-week level. Every observation corresponds to one SKU in one calendar week and includes the weekly sales amount together with indicators for special cases, such as holiday weeks and return-related negative entries. This is done to preserve the temporal ordering of demand, allowing direct aggregation from SKU level to total sales and supporting both descriptive analysis and forecasting on a common weekly time grid.

In order to provide a clearer view of the company's sales development over time total annual sales are presented in Figure 3.2. This figure gives an overview of the overall sales volume during the period 2022–2025 and helps contextualise the size of the dataset before moving to SKU-level analysis.

In addition to annual totals, the weekly sales series is shown in Figure 3.3. The weekly view is important because it reveals short-term fluctuations and large spikes between weeks, which are not visible in annual summaries. These weekly movements are directly relevant to the forecasting task, since the model is intended to support short-term planning decisions.

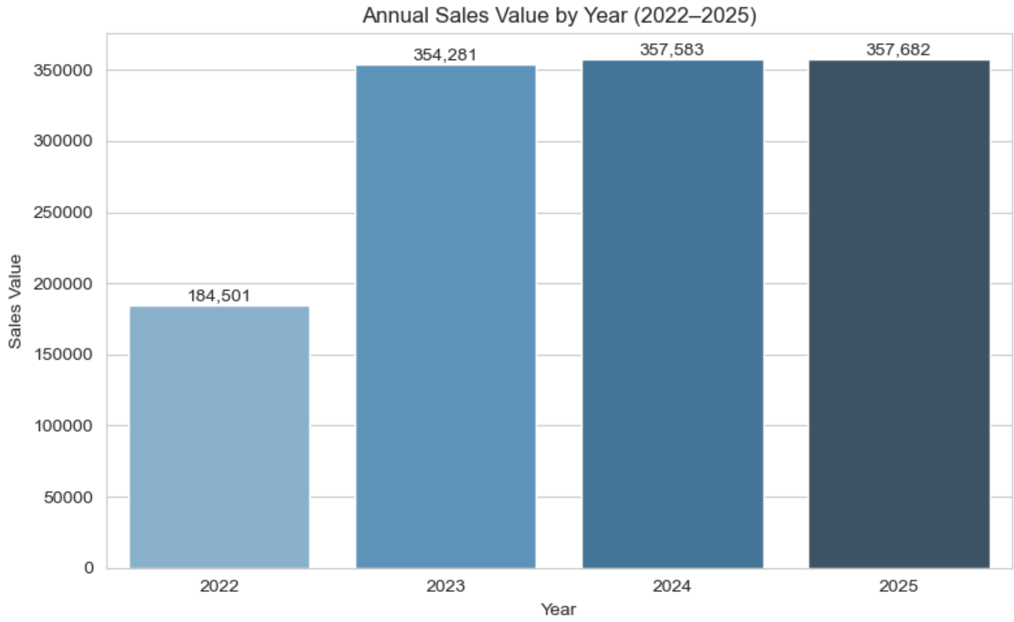

Figure 3.2: Total annual sales for the period 2022–2025

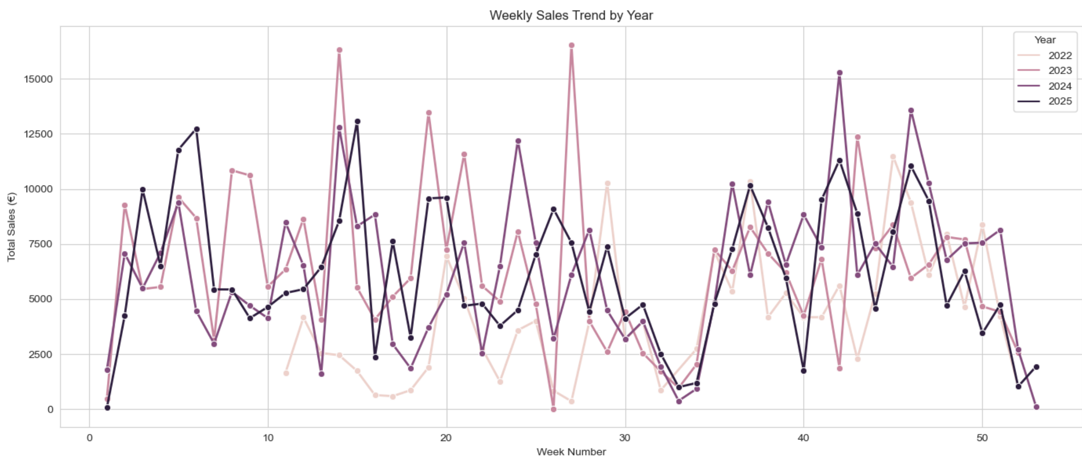

Figure 3.3: Weekly sales series, 2022–2025

3.4.3 Handling Missing Values and Zero Sales

A central issue in the preprocessing stage was the distinction between missing observations and genuine zero-sales weeks. In the final weekly panel, weeks with no recorded sales were retained as zero-demand observations rather than being treated as missing values. This distinction is important because a zero-sales week may indicate that no demand has happened and a missing value would indicate that information is unavailable. Treating these cases incorrectly could distort the demand pattern and affect the choice of forecasting method.

The summary statistics in Table 3.2 show that demand is highly uneven across SKUs. The median mean weekly demand is only 6.618, while the maximum reaches 313.500. This shows strong right-skewness, meaning that most SKUs have relatively low weekly demand and a small number of products account for much higher sales volumes.

Table 3.2: Summary statistics of SKU-level demand measures

Statistic	Mean Weekly Demand	Std Weekly Demand	Arrival Rate	Mean Case Demand
Count	448.000	448.000	448.000	448.000
Mean	20.990	27.085	0.344	41.585
Std. Dev.	39.484	46.862	0.325	69.693
Min	0	0	0	0
25%	0	0	0	0
Median	6.618	11.898	0.262	21.100
75%	22.676	32.452	0.592	47.743
Max	313.500	408.762	1.000	606.797

The arrival rate has a mean of 0.344 and a median of 0.262 and to understand that for a typical SKU, positive sales occur in only about a quarter to a third of the observed weeks. This confirms that many products exhibit intermittent demand rather than stable weekly demand. The mean case demand is higher than the mean weekly demand because it is calculated only over weeks with positive sales. Its median value of 21.100 suggests that when demand occurs, it is often moderate, while the maximum value of 606.797 shows that a few SKUs experience very large demand spikes.

Figure 3.4 presents the distribution of mean weekly demand per SKU on a log scale. The histogram shows that most SKUs are concentrated at low demand values and only a few products reach very high demand levels. The log scale is useful because it makes the full distribution visible without compressing the low-demand SKUs.

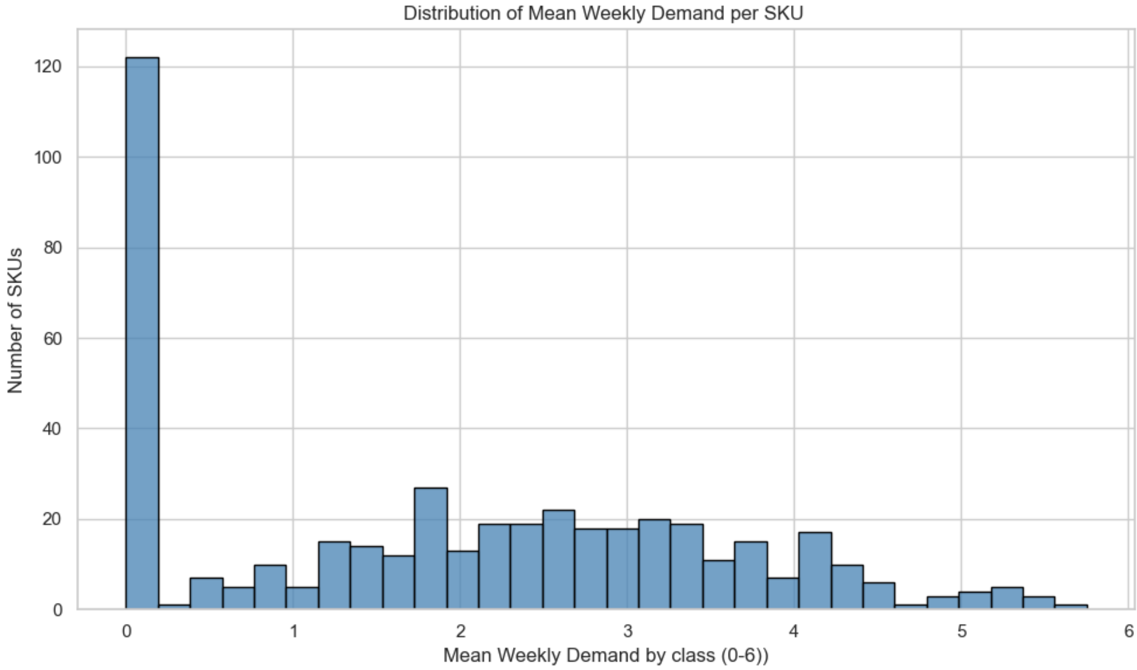

Figure 3.4: Distribution of mean weekly demand per SKU on a log scale

Figure 3.5 shows the distribution of the arrival rate per SKU. The concentration of values near zero indicates that many products have long stretches of zero sales.This is promising because this pattern is typical of intermittent demand and supports the use of forecasting methods designed for sparse series, such as Croston-type methods.

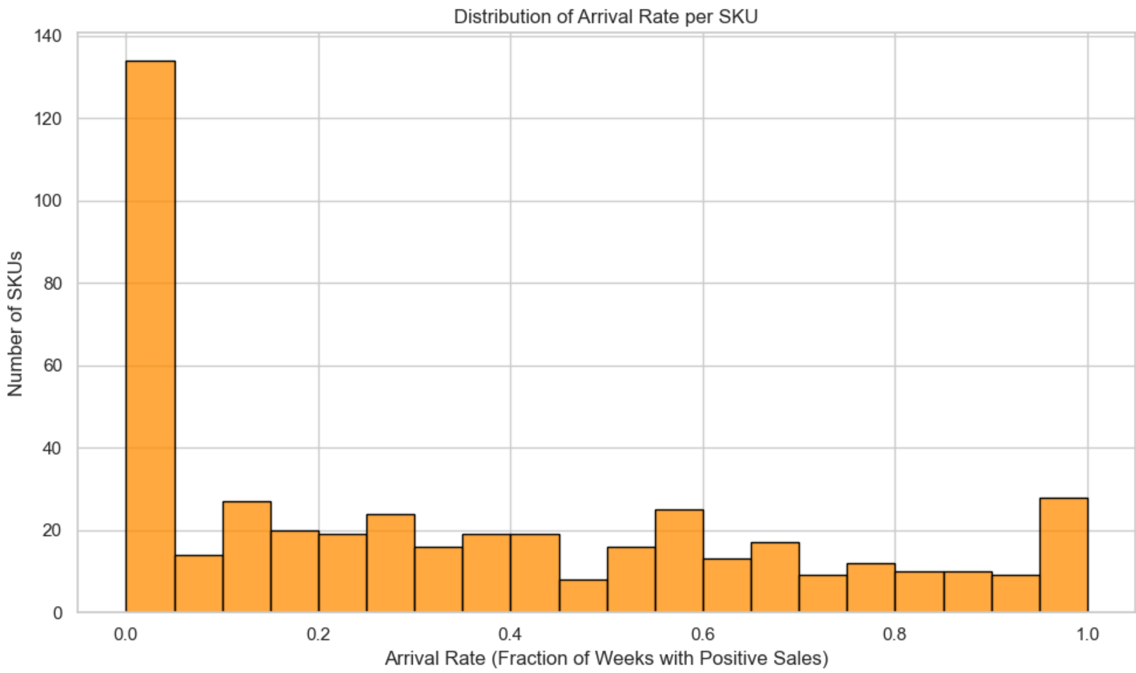

Figure 3.5: Distribution of arrival rate per SKU

Finally, Figure 3.6 compares mean weekly demand with the standard deviation of weekly demand for each SKU. The scatter plot shows that variability tends to rise as average demand increases which means that high-demand SKUs are not only more important in terms of volume, but also more volatile. Thus, they may require different treatment from low-volume or highly intermittent products during forecasting and segmentation.

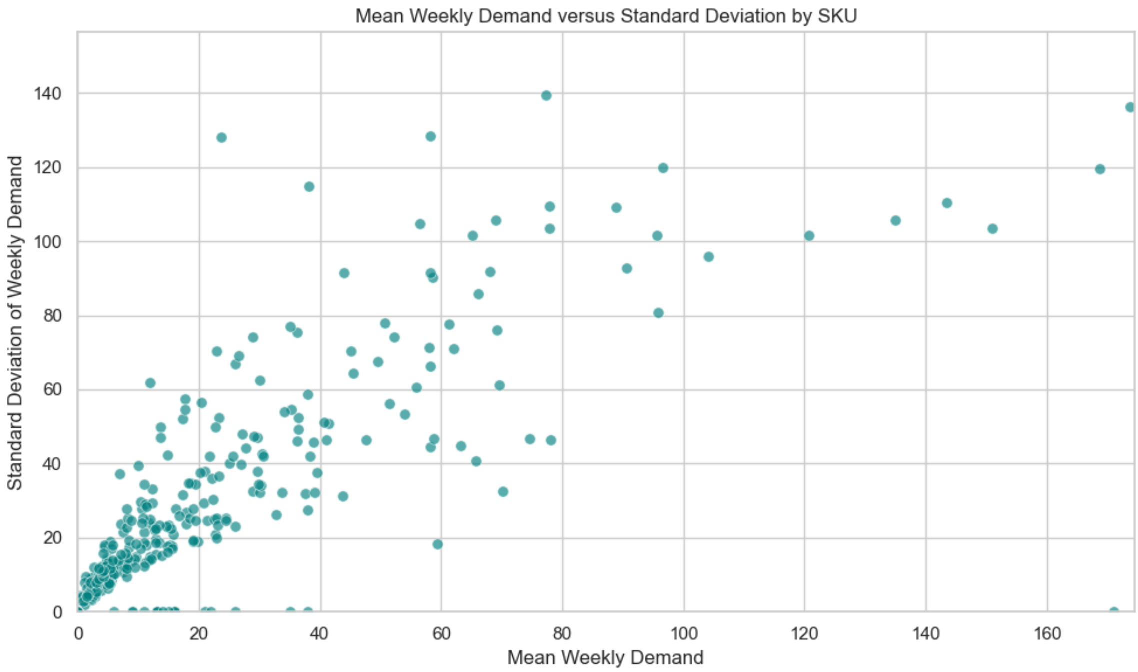

Figure 3.6: Mean weekly demand versus standard deviation by SKU

To sum up, the preprocessing procedure produced a cleaned and standardised weekly dataset that preserves the main characteristics of the original sales history. Zero-sales weeks were retained as meaningful observations, negative or non-standard entries were flagged, and the final structure supports both SKU-level forecasting and aggregation to higher levels where needed. This dataset forms the basis for the exploratory analysis, segmentation, and forecasting experiments presented in the following sections.

3.5 Exploratory Data Overview

3.5.1 Weekly sales trajectories of selected SKUs

However, A picture is worth a thousand words and to examine products with relatively limited zero-sales incidence in greater detail, three representative SKUs are selected and their weekly sales trajectories are plotted. The selected items are SKU 31710007 (*Gift Box Dark Cherry*), SKU 31300286 (*Eau De Toilette Dark Cherry*), and SKU 31310001 (*Eau De Toilette Golden Honey & Argan Oil*), all belonging to the *Blue Scents* brand family.

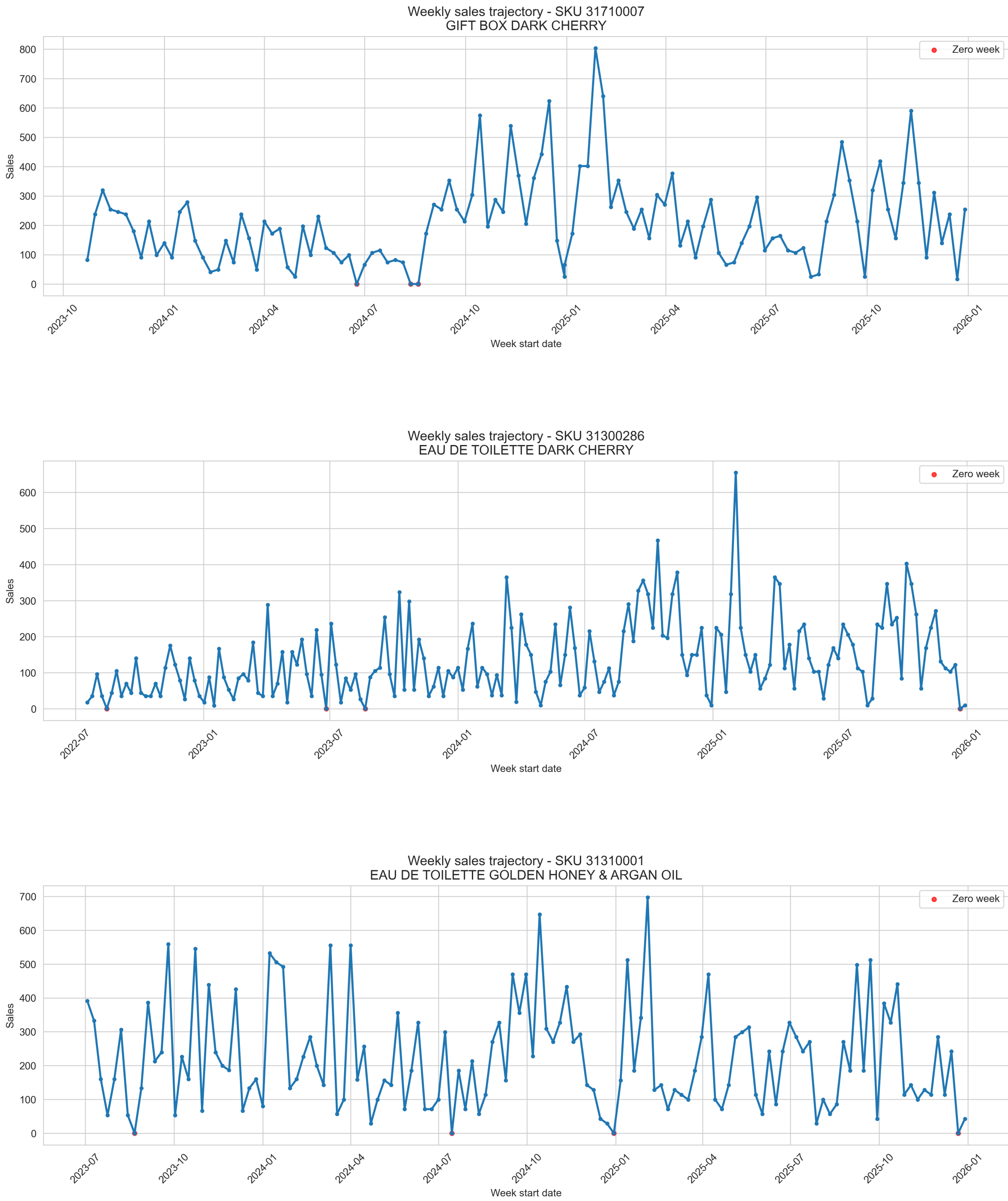

Figure 3.7: Weekly sales trajectories for the selected SKUs 31710007, 31300286, and 31310001

As shown in Figure 3.7, the weekly sales trends of the three chosen SKUs are not smooth and differ significantly from week to week. Although zero-sales weeks are relatively few in these products, the series still feature distinct peaks, sudden troughs, and short-term fluctuations suggesting that demand does not evolve in a stable manner over time. SKU 31710007 (*Gift Box Dark Cherry*) appears to be the most volatile of the three series with steep upward spikes followed by observable corrections. SKU 31300286 (*Eau De Toilette Dark Cherry*) shows a similar pattern with local maxima and minima that repeat over the observed time span. SKU

31310001 (*Eau De Toilette Golden Honey & Argan Oil*) appears somewhat more consistent but still fluctuates significantly in its weekly sales and frequently behaves abruptly.

This is consistent with the fact that demand may be affected not only by a static flow of baseline products, but also by short-term commercial interventions. Discounts, bundle offers, campaign periods, or similar promotional activities generate short-term boosts in sales and can explain some of the spikes seen in the plots. It is therefore useful to perform a follow-on analysis based on promotion information related to the recorded promo codes and linked to the source data.

Table 3.3: Selected SKUs included in the weekly trajectory analysis

SKU	Description	Brand	Supplier	Category	Subtype
31710007	GIFT BOX DARK CHERRY	BLUE SCENTS	BLUE SCENTS	GIFT BOX	DARK CHERRY
31300286	EAU DE TOILETTE DARK CHERRY	BLUE SCENTS	BLUE SCENTS	EAU DE TOILETTE	DARK CHERRY
31310001	EAU DE TOILETTE GOLDEN HONEY & ARGAN OIL	BLUE SCENTS	BLUE SCENTS	EAU DE TOILETTE	ARGAN OIL GOLDEN HONEY

3.5.2 ABC–XYZ segmentation and model mapping

Following the cleaning and preparation of the weekly sales dataset, an ABC–XYZ segmentation is performed in order to classify products according to both sales importance and demand behaviour. The ABC classification is based on sales contribution variables such as total sales, sales share, and cumulative sales share. ”A” means items contribute the most, ”B” means that items are intermediate and ”C” means items contribute the least. The XYZ classification captures demand regularity using indicators such as non-zero weeks, zero weeks, and the coefficient of variation (CV). ”X” means that items have stable demand, ”Y” means items show moderate fluctuations and ”Z” means that items have highly irregular or unpredictable demand.

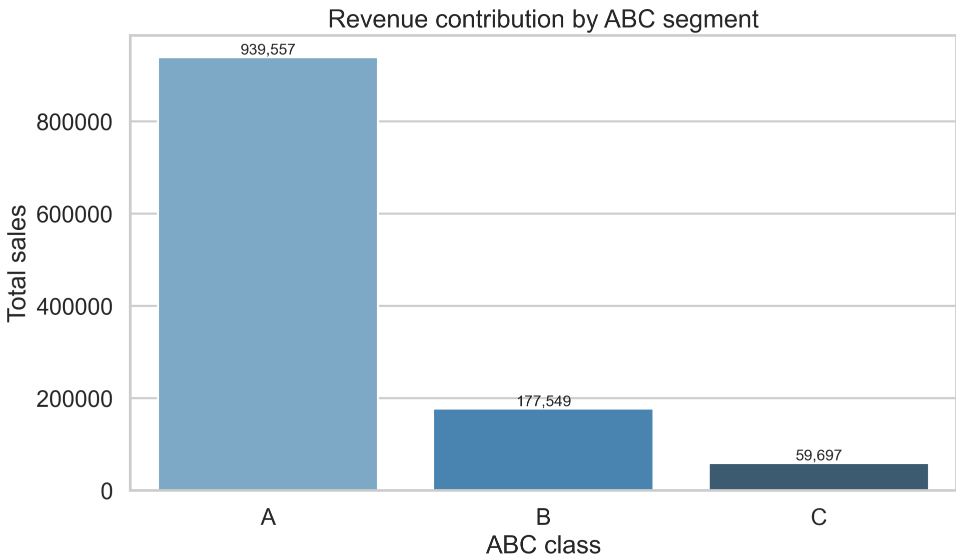

Figure 3.8: Revenue contribution by ABC segment

Figure 3.8 shows the revenue contribution of the three ABC classes. Class A products generated total sales of 939,557, compared with 177,549 for class B and 59,697 for class C. This indicates a highly concentrated sales structure, in which a relatively small subset of SKUs accounts for the majority of revenue.

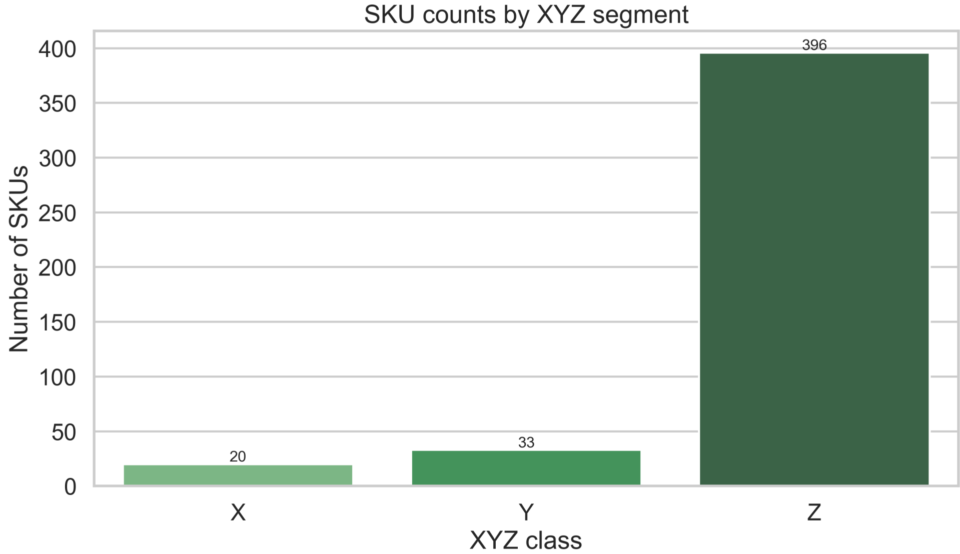

Figure 3.9: SKU counts by XYZ segment

Figure 3.9 presents the number of SKUs assigned to each XYZ class. The distribution is heavily concentrated in class Z, which contains 396 SKUs, while only 33 SKUs belong to class Y and 20 to class X. This pattern suggests that most products exhibit irregular or intermittent weekly demand, whereas only a limited number show relatively stable demand behaviour.

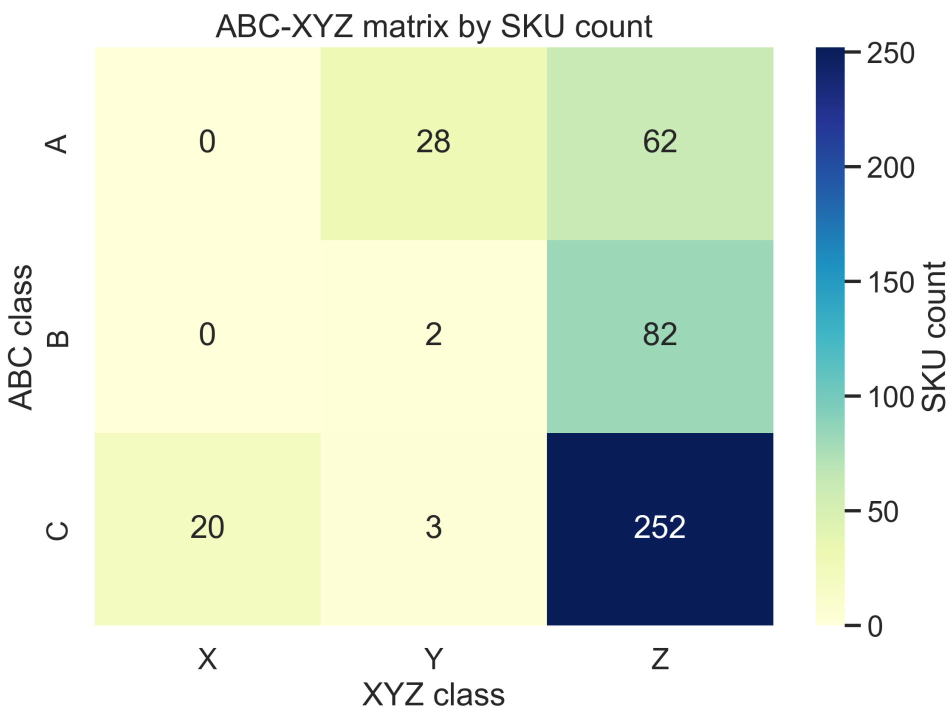

Figure 3.10: ABC–XYZ matrix by SKU count

Figure 3.10 presents the ABC–XYZ matrix in terms of SKU counts. The largest concentration of items is found in segment CZ with 252 SKUs, followed by BZ with 82 and AZ with 62, while no A or B items are classified in class X. This shows that irregular demand is widespread across the assortment and is not limited to low-priority products.

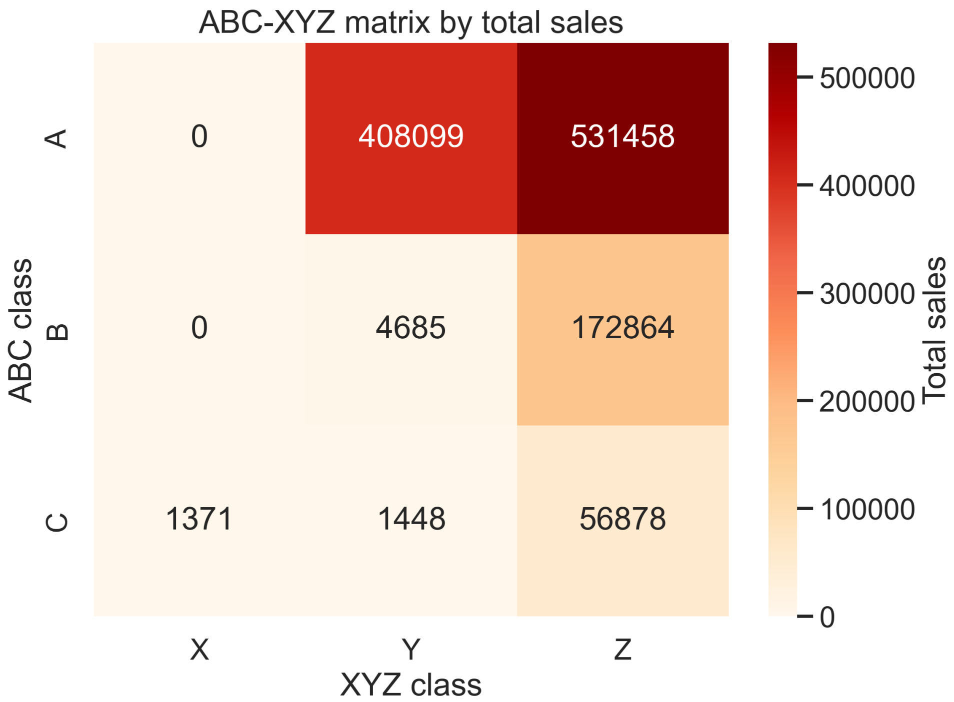

Figure 3.11: ABC–XYZ matrix by total sales

Figure 3.11 presents the ABC–XYZ matrix in terms of total sales. The largest revenue contribution is generated by AZ products, with total sales of 531,458, followed by AY products with 408,099. This is an important result because it shows that a substantial share of commercial value is associated with irregular demand patterns. Lastly,the segmentation results indicate that the product portfolio is concentrated in revenue terms but dominated by irregular demand in behavioural terms. The ABC–XYZ framework therefore provided a useful preprocessing step for prioritising SKUs by business importance and assigning forecasting models according to empirical demand characteristics.

Chapter 4

Feature Analysis and Selection

4.1 Feature Analysis and Selection

This chapter examines the main explanatory features considered for the forecasting framework and the purpose of the feature analysis is not only to create additional variables, but also to understand which demand drivers are meaningful, interpretable and suitable for short-term SKU-level forecasting. Since the literature review highlighted the importance of combining historical demand patterns with event-related information, this chapter focuses on promotion variables, holiday and calendar effects, lagged demand information, price-related indicators, and ABC–XYZ segmentation. These features are then assessed in terms of their practical relevance, data availability, and suitability for the final modelling framework.

4.1.1 Feature Engineering Overview

Feature engineering is an essential step in the forecasting framework as observed sales data on their own do not fully explain short-term demand variation at SKU level. Demand may be affected by seasonality or promotions or holiday effects or price-related changes or even product-specific behaviour, all of which can alter the observed sales pattern. The role of feature construction is to convert these sources of information into structured variables that can be used consistently across the forecasting pipeline.

During the analysis the feature set is developed around promotion indicators, calendar-based variables, lagged demand features, and price-related proxies. All variables are defined at weekly frequency to match the structure of the target series and the intended forecasting horizon. In addition the ABC–XYZ segmentation which mentioned in the data chapter is also used to support the later model selection process.

The resulting design aims to balance predictive usefulness with operational interpretability rather than retaining every potentially available variable and each feature is assessed according to whether it can reasonably explain short-term demand changes, whether it is known at the time of forecasting, and whether it contributes to the practical usefulness of the final forecasting system.

4.1.2 Promotion Features

Promotions are a central feature group in Just Health because they represent planned commercial events that may temporarily alter demand. In the forecasting literature, promotional activity is commonly treated as an external demand driver because it can create short-term sales spikes that are not fully captured by historical sales patterns alone (Abolghasemi et al. 2020; Fildes et al. 2022). The objective of the promotion feature design is therefore to convert the company's raw planning information into variables that can be used directly in the forecasting framework.

The raw promotion file contains campaign-level observations, where each promotion is identified by SKU, promotion type, start date, end date, and duration.The construction of this file, has been done by analytics department by transforming raw data into a large database. These campaign records are then transformed into a weekly SKU-level structure, so that each SKU-week observation reflects whether a promotion is active and, if so, what type of promotional activity is present. Furthermore, the feature set captures timing and duration effects allowing the model to distinguish between isolated promotional weeks and longer or more structured campaigns.

Table 4.1: Promotional and Holiday Coverage by Demand Segment

Segment	**Promo Weeks**	**Promo Ratio**	**Holiday Weeks**	**Holiday Ratio**
AY	639	0.132	907	0.188
AZ	696	0.073	1,805	0.189
BY	21	0.114	34	0.184
BZ	497	0.042	2,202	0.187
CX	3	0.094	2	0.062
CY	1	0.050	3	0.150
CZ	78	0.004	4,134	0.189
Total	**1,935**	**0.040**	**9,087**	**0.188**

Note: Promo Ratio and Holiday Ratio are computed as the proportion of weekly observations within each segment that contain at least one active promotional event or a Greek public holiday, respectively.

4.1.2.1 Construction of Promotion Variables

The original promotion planning dataset contains the variables `SKU`, `Promotion`, `Start Date`, `End Date`, and `Duration weeks`. This file defines the event structure of the promotional calendar. A preview of the promotion planning dataset is shown in Table 4.2.

Table 4.2: Preview of the promotion plan dataset

SKU	Promotion	Start Date - End Date	Duration (weeks)
13000009	Promotion stand	20/02/2023–05/03/2023	2
13000009	Flyers	08/05/2023–14/05/2023	1
13100001	Price-off / Bundle	16/05/2022–22/05/2022	1
13100001	Promotion stand	29/08/2022–18/09/2022	3
13100001	1+1	21/11/2022–27/11/2022	1

The construction process can be described in four stages. First, the raw promotion plan is used to identify all campaigns and their timing. Secondly, each campaign is mapped into the weekly calendar of each SKU, so that a multi-week campaign appears in every affected week. Thirdly, promotion-type dummy variables are created to distinguish between Flyers, Price-off / Bundle, 1+1 promotions, and Promotion Stand campaigns. Fourthly, intensity and recency variables are engineered to capture not only whether a promotion occurred, but also how strong the promotion was and how recently another promotion had taken place.

The resulting weekly promotion feature file includes a binary indicator of promotional activity, a weekly promotion count, campaign duration in weeks, sales uplift measures, the peak turnover reached during the promotional period, one-hot encoded promotion-type indicators (`flyers`, `priceoffbundle`, `promo11`, `promotionstand`), a composite promotion intensity score, and temporal recency features. These variables move the analysis beyond a simple yes/no treatment of promotions and allow the model to capture heterogeneous effects across timing, duration, and promotion mechanism.

This transformation is important because the forecasting models operate at weekly SKU level, while the original promotion information is stored at campaign level. By expanding the campaign information into the same time grid as the sales data, promotion effects can be analysed together with weekly demand and can later be used as explanatory variables in promo-aware forecasting models.

4.1.2.2 Promotion Frequency and Timing Patterns

The first descriptive step is to examine how promotion activity is distributed across SKUs. Figure 4.1 presents the distribution of promotion counts across the assortment.

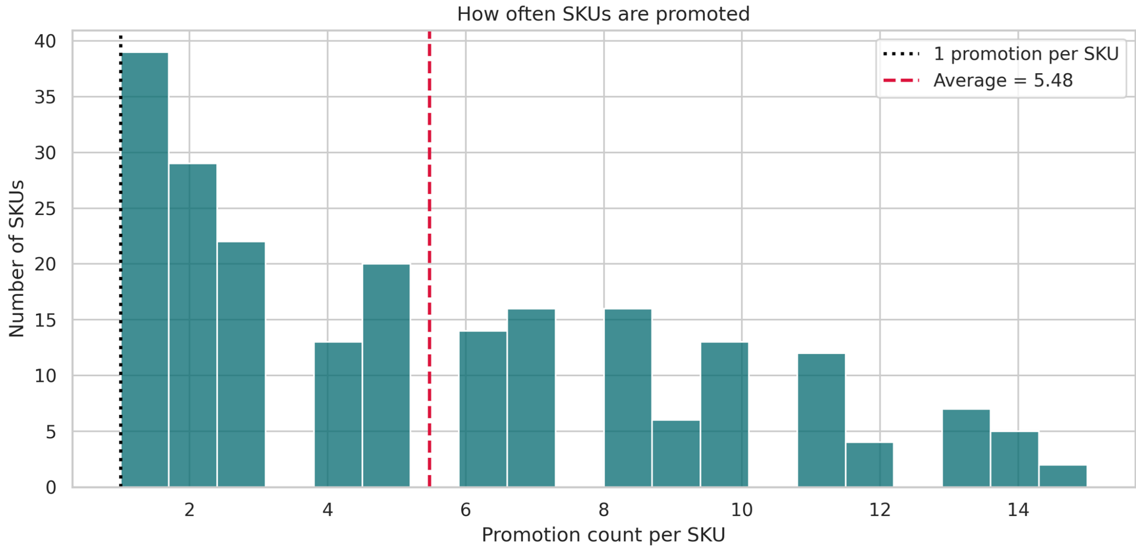

Figure 4.1: Promotion frequency per SKU.

If someone analyze Figure 4.1 it is understanble that promotion activity is unevenly distributed across the product portfolio. Most SKUs appear in promotions only a few times, while a smaller number of SKUs are promoted more frequently. This is happening because certain companies promote certain products more often and it is not pharmacy's choice which and when a promotion will happen. The dashed reference line marks one promotion per SKU.This provides a useful benchmark for interpretation. Based on this, a total of about 400 promotions would correspond to roughly one promotion event for each SKU in an assortment of 448 SKUs. This helps contextualise the intensity of promotional activity relative to the size of the product assortment.

The promotion mix is then examined by campaign type. Figure 4.2 shows the number of promotion events by type.

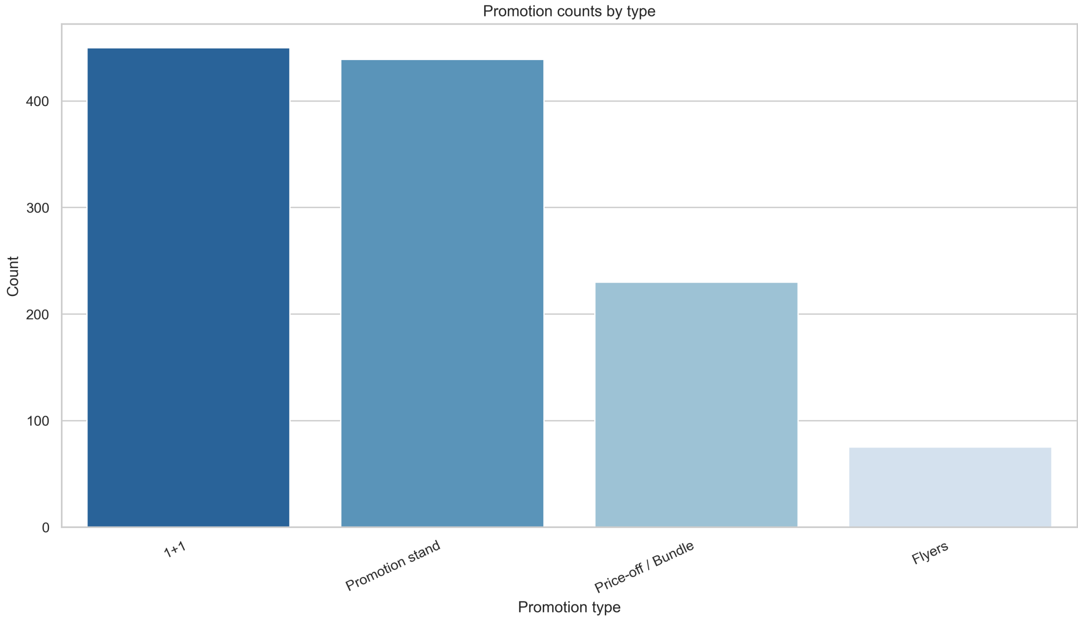

Figure 4.2: Promotion counts by type

Figure 4.2 indicates promotion mix is dominated by `1+1` and `Promotion Stand` campaigns, which appear considerably more frequently than `Price-off / Bundle` promotions and particularly `Flyers`. This suggests that the promotional strategy in the observed portfolio relies primarily on repeated in-store support and multi-buy mechanics rather than on flyer-led communication. Since `1+1` and `Promotion Stand` campaigns account for most promotional weeks, they are also expected to be the strongest statistical drivers of the promotion-demand relationship. This led to decide that it is needed to encode promotion categories separately rather than collapsing them into a single generic promotion flag.

The timing of promotions is also relevant, because promotional activity may be clustered in specific periods instead of being evenly distributed over time. Figure 4.3 presents the monthly frequency of promotions across the sample period.

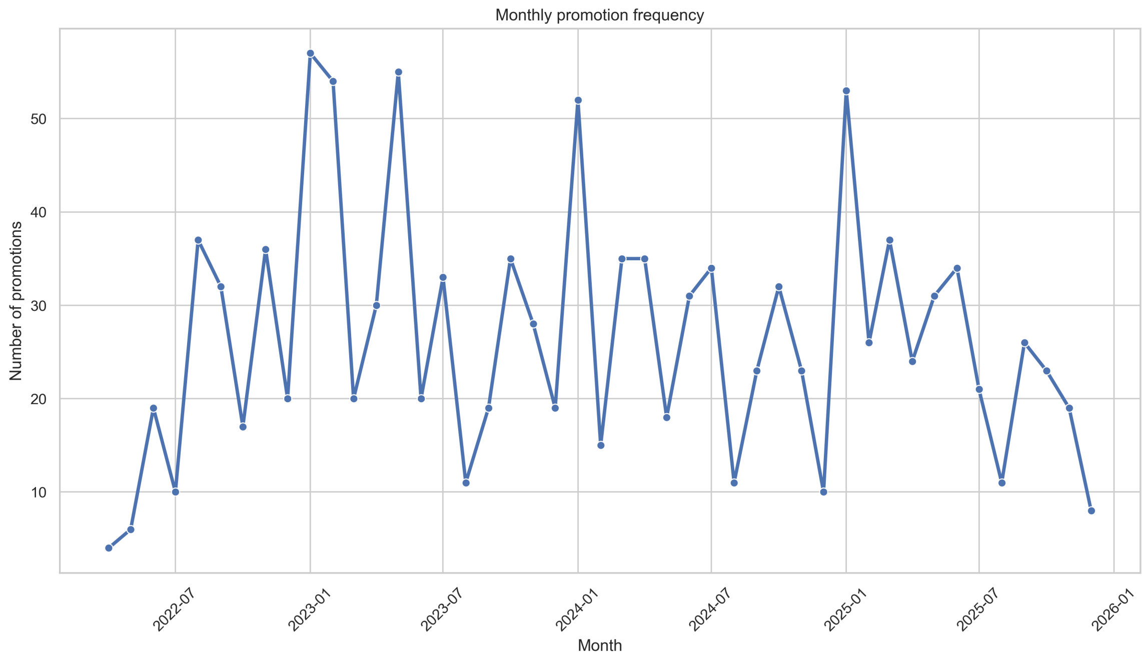

Figure 4.3: Monthly promotion frequency

Figure 4.3 shows that promotional activity is not constant over time, for example , several months exhibit clear peaks whereas others show comparatively low activity. This indicates that the commercial calendar is clustered rather than uniform. Such clustering is important for forecasting because it means that promotion effects may coincide with seasonal periods, holiday periods, or broader commercial planning cycles.

The duration of promotion campaigns is analysed in Figure 4.4. Duration is relevant because short and long campaigns may influence demand differently. A one-week campaign may create a concentrated spike, while a longer campaign may spread the effect across several weeks.

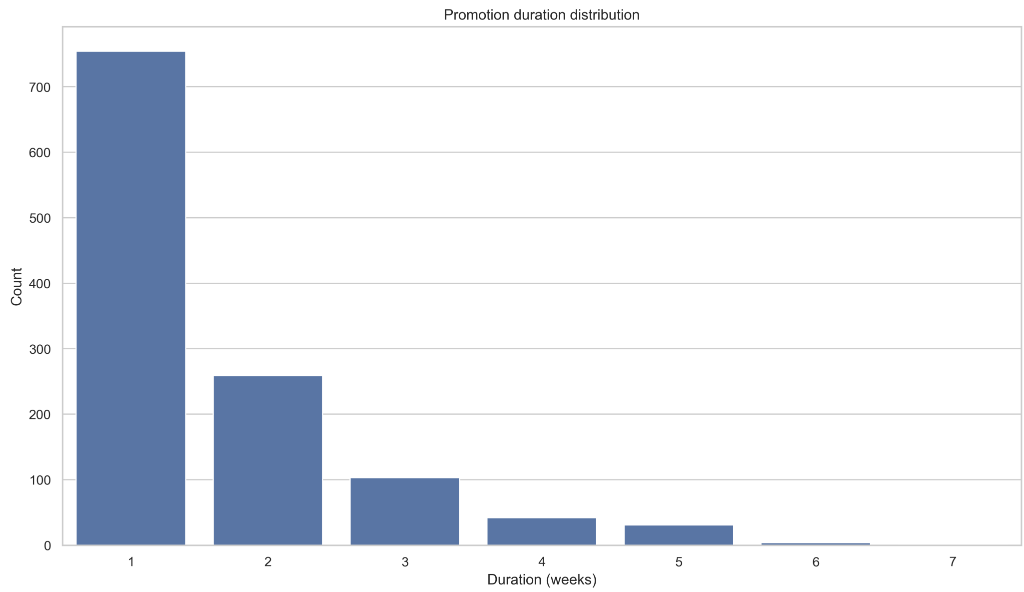

Figure 4.4: Promotion duration distribution

To explain that, the duration profile is heavily concentrated in one-week promotions, with a steep decline as duration increases. Two-week promotions still appear with meaningful frequency, but campaigns longer than three weeks are much rarer. This indicates that the promotional structure is mainly driven by short tactical interventions rather than long-running campaigns.

Because short promotions dominate the sample, duration is retained as an informative descriptive variable. Short campaigns may create brief and concentrated demand effects, while longer campaigns may reflect more sustained commercial actions. For this reason, campaign duration is useful not only for descriptive analysis but also for interpreting the expected timing of promotion-driven sales changes.

The spacing between promotions is examined in Figure 4.5. This variable captures the number of weeks since the previous promotion and helps describe whether promotional activity occurs in repeated bursts or with long gaps between campaigns.

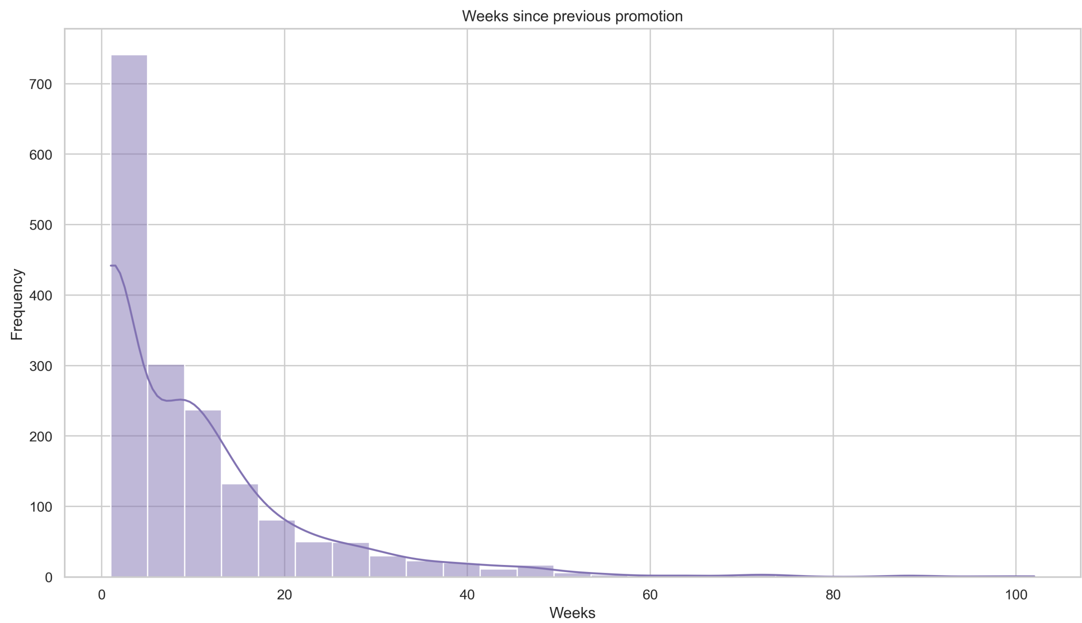

Figure 4.5: Weeks since previous promotion

Figure 4.5 help us understand that the spacing between promotions is uneven. Some products are promoted again after relatively short gaps, while others remain without promotion for longer periods. This pattern indicates that promotional activity is not applied uniformly across the assortment, but follows product-specific timing and commercial priorities.

Taken together, the frequency, type, duration, and spacing patterns show that promotions are concentrated in a non-random way both over time and across products. This is useful for understanding the commercial calendar of the dataset, since it highlights that promotional activity is structured rather than evenly distributed.

4.1.2.3 Interpretation of Promotion Effects

In general, the promotion analysis shows that the promotions are heterogeneous and unevenly distributed between the types of campaigns and SKUs. This means that treating promotions as an important feature group in the forecasting framework is vital. If promotions were ignored, the model could misinterpret promotion-driven spikes as random demand variation or as part of the normal time-series pattern.

The descriptive comparison between promotional and non-promotional weeks indicates that promotional activity is associated with substantially higher weekly sales values. In addition, the variation across promotion types suggests that the effect is not uniform, but depends on the specific promotion mechanism applied. This reinforces the need to model promotions as heterogeneous explanatory variables rather than as a single binary signal.

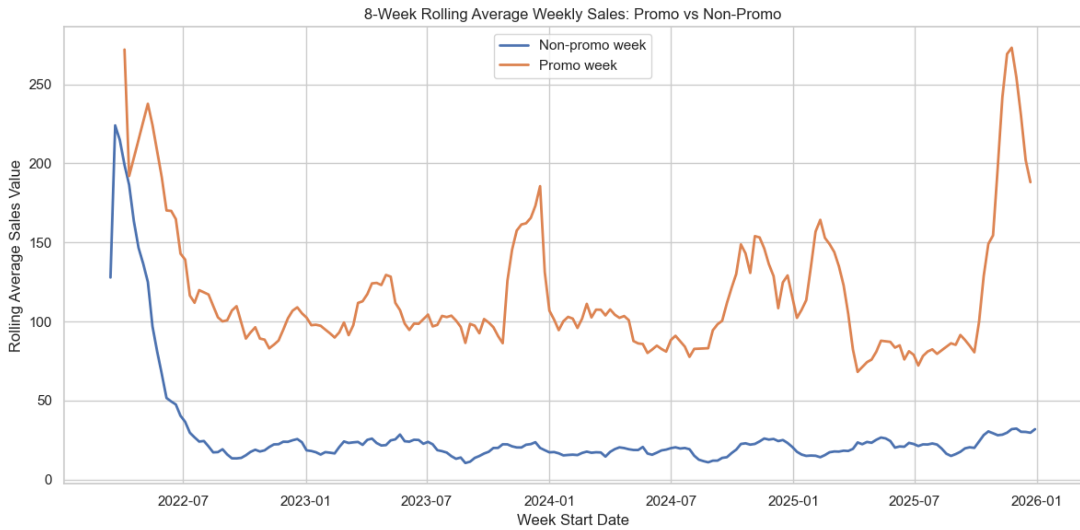

Figure 4.6: Descriptive comparison of promotion-related demand effects

Figure 4.6 supports the broader feature-engineering strategy adopted in this study, where raw campaign information is transformed into forecasting-ready variables that preserve timing and duration information. The role of promotion features is therefore not to replace the core demand history, but to complement it with operational information that can explain short-run deviations from the baseline sales pattern.

At the same time, promotion variables should be used carefully. Not every SKU has the same promotion history, and some products may have too few promotion observations to estimate a reliable effect. Therefore, promotion features are most useful for SKUs with sufficient historical promotional activity, while simpler univariate benchmarks remain important for SKUs with limited external-event information.

4.1.3 Holiday and Calendar Features

In addition to promotions, calendar effects related to public holidays are examined in order to assess whether they contribute to systematic changes in weekly demand because it is needed to have another indicator in order to have better results. Holiday effects are relevant because demand may change around public holidays due to altered consumer routines, store operations, logistics schedules, or shifts in weekly purchasing behaviour. As discussed in the literature review, calendar variables can improve interpretability when they help separate normal demand variation from event-driven deviations (Huber and Stuckenschmidt 2020; Hyndman and Athanasopoulos 2021).

The purpose of this analysis is both descriptive and methodological. From a descriptive perspective, it helps identify whether specific holidays are associated with higher or lower sales and from a methodological perspective, it helps decide whether holiday variables should be incorporated into the final forecasting framework as explanatory variables. A preview of the weekly holiday impact dataset is shown in Table 4.3.

Table 4.3: Preview of the weekly holiday impact dataset

Date	Holiday	Week start	Year	Week	Sales	Uplift vs median (%)
2022-05-01	Labour Day (Protomagia)	2022-04-25	2022	17	593	-85.75
2022-06-13	Holy Spirit Monday	2022-06-13	2022	24	3630	-12.77

The holiday dataset aligns each holiday with its corresponding week start date, year, week number, observed sales, and uplift relative to a median benchmark. This structure allows the analysis to compare holiday weeks with normal sales behaviour while preserving the weekly time scale used in the forecasting task.

4.1.3.1 Week-long Holiday Effects

The weekly holiday analysis compares sales in holiday weeks with the annual median benchmark and uses the resulting difference as a descriptive uplift measure for each holiday instance. This approach is aligned with the forecasting horizon because the models are designed to operate at weekly frequency. As a result, the weekly view is the most directly relevant perspective for deciding whether holiday indicators should be included in the forecasting framework.

Positive average weekly uplift is reported for only a small number of holidays. Clean Monday delivered an average weekly uplift of 22.88%, whereas Ochi Day recorded an average uplift of 18.71%. Greek Independence Day presents a more ambiguous pattern, with an average uplift of -7.75% and a positive median uplift of 5.86%. Because only a limited number of observations are available for some individual holidays and the reported percentages are relatively small, these results should be interpreted cautiously. Overall, the pattern suggests that Greek Independence Day does not exert a uniformly negative influence, but rather a mixed effect.

In contrast, most holidays are associated with strongly negative weekly performance. The Dormition of the Virgin Mary generated the largest negative average weekly uplift at -87.60%, followed by New Year at -82.30%, Boxing Day at -69.02%, Epiphany at -64.14%, and Orthodox Easter Day and Orthodox Good Friday at -56.82%. These results indicate that many holidays correspond to suppressed weekly demand rather than demand peaks. This means that during holidays people tend not to go to pharmacies so often.

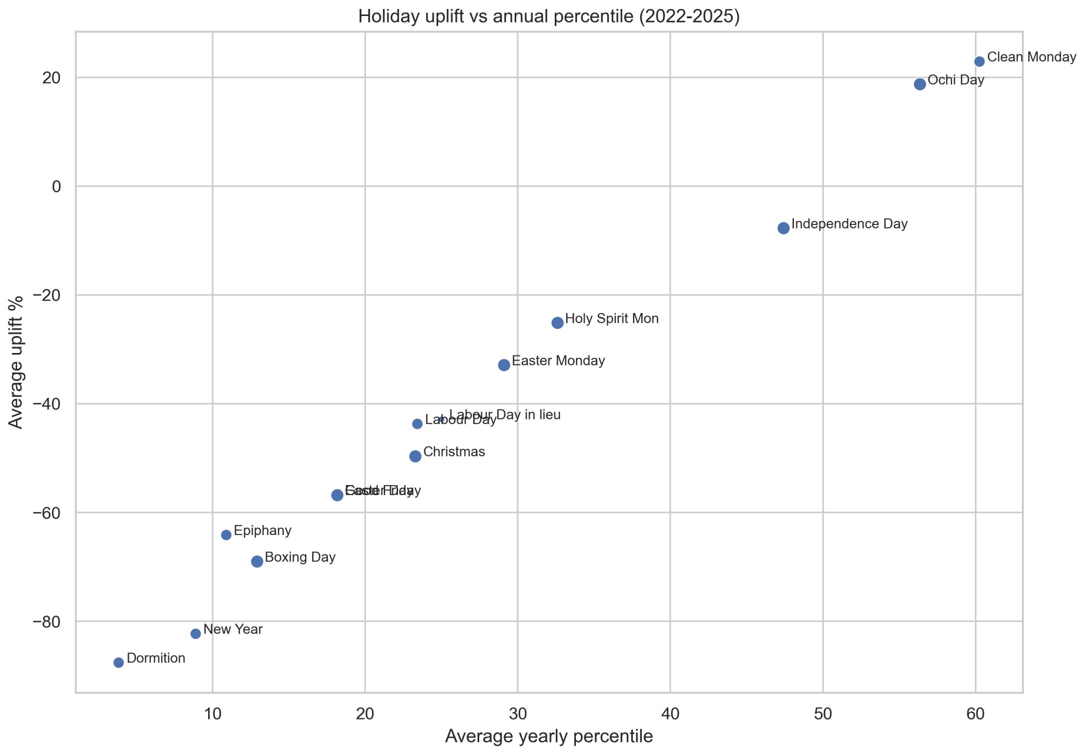

Figure 4.7: Average weekly uplift by holiday relative to the annual median benchmark

Figure 4.7 shows that holiday effects are not uniform. Some holidays are associated with positive or mixed demand effects, while others are clearly linked to lower weekly sales. This finding is important because it suggests that a single binary variable indicating whether a week contains any holiday would be too coarse. Such a variable would combine positive, neutral, and negative holidays into the same category and could therefore obscure the actual relationship between holidays and demand.

4.1.3.2 Monthly Holiday Effects

A complementary monthly analysis is used to examine whether the holiday-related weekly patterns remain visible when sales are aggregated over the month in which the holiday occurs. This aggregation is more suitable than comparing a single holiday day in isolation, because the goal is to capture the broader month-level sales context in which the holiday is embedded. Despite the fact that this approach is still relatively coarse, it helps indicate whether a holiday is associated only with its own week or with a wider monthly sales effect.

At the monthly level, Ochi Day remains the strongest positive holiday effect, with an average uplift of 23.41%. Labour Day also appears positive, with an average uplift of 10.04%, while Epiphany and New Year both record average monthly uplifts of 3.78%. In contrast, Dormition remains strongly negative even after aggregation, with an average monthly uplift of -54.50%. Clean Monday, which is positive on average in the weekly analysis, becomes negative at the monthly level with an average uplift of -27.45%, showing that its effect is sensitive to the level of temporal aggregation.

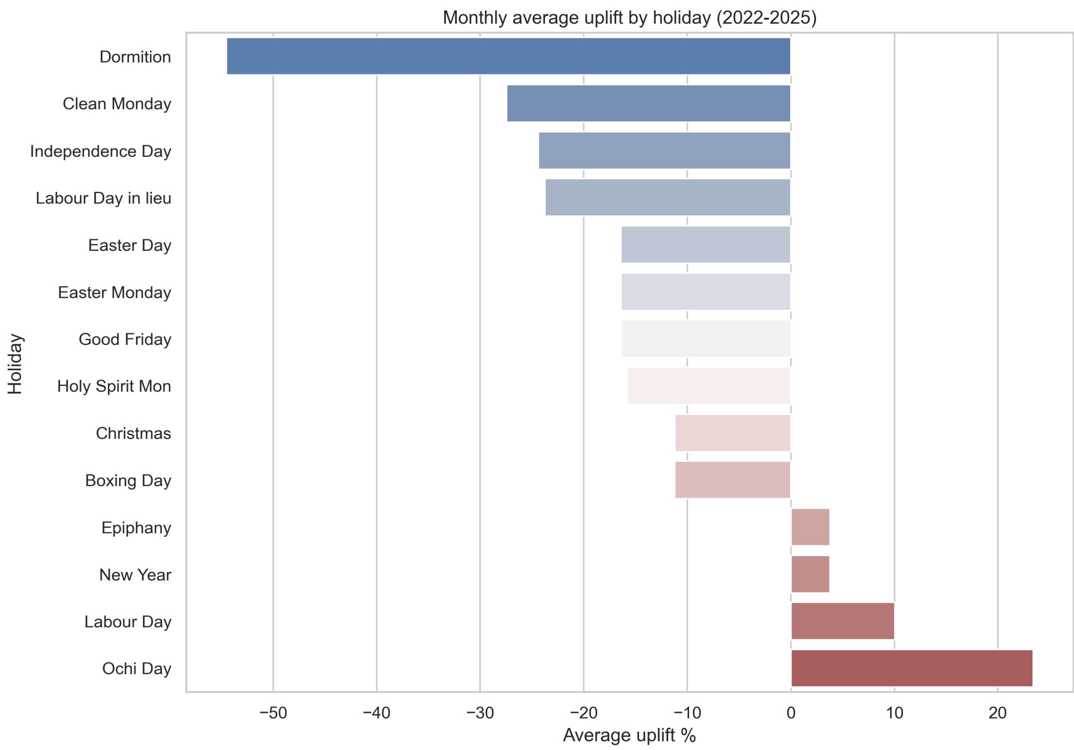

Figure 4.8: Average monthly uplift by holiday relative to the monthly median benchmark

Since the forecasting framework is weekly, the weekly holiday interpretation should be prioritised because it is more directly aligned with the prediction target. The monthly analysis nevertheless remains useful as a robustness check as it shows that some holiday effects remain visible after aggregation, while others change direction depending on the temporal level of analysis. Moreover it is needed to evaluate holiday effects carefully rather than assuming that all holidays influence demand in the same way.

4.1.3.3 Variation Across Years

The year-by-year analysis shows that holiday effects vary substantially across years and are not stable over time. Some holidays show pronounced year-to-year differences suggesting that their sales impact reflects broader commercial and calendar conditions rather than a fixed deterministic effect. This is important for forecasting because a holiday indicator based on only a few observations may not generalise reliably across years.

Ochi Day is the clearest example of this instability, ranging from -43.79% in 2022 to 106.21% in 2023, then 37.26% in 2024, and -24.83% in 2025. This variation may partly reflect differences in the day of the week on which the holiday falls, as well as interactions with surrounding seasonal and promotional conditions. Clean Monday shows a similarly uneven pattern, moving from a strong positive effect in 2023 to a near-neutral effect in 2024 and a negative effect in 2025.

However, some holidays are more consistently negative across years.For example, Dormition still remains negative in all years, while Christmas, Boxing Day, New Year, and the Easter-related holidays also tend to suppress sales. This distinction between unstable and consistently negative holidays is useful for feature selection, because it suggests that some holiday effects may be modelled separately and others may be grouped as disruption-related calendar controls.

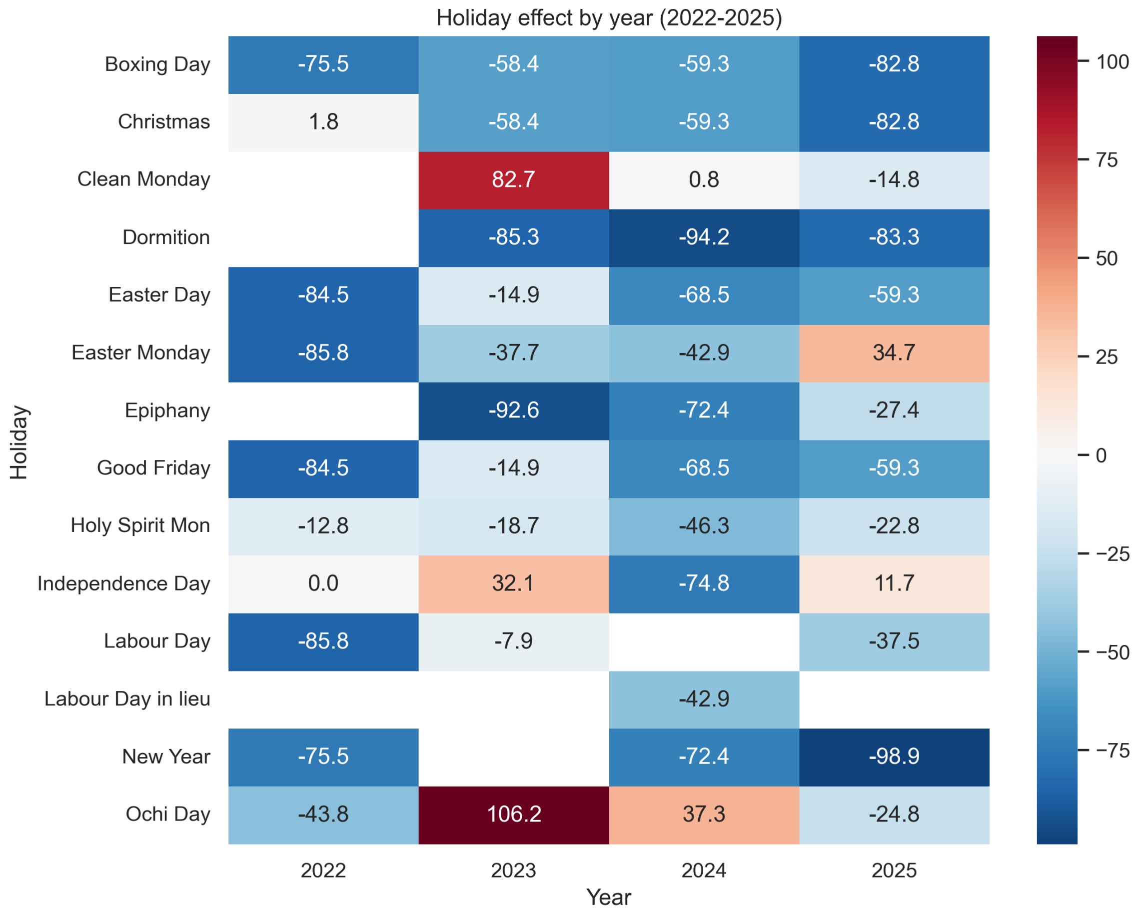

Figure 4.9: Holiday uplift by year

Figure 4.9 summarises this variation across years. The heatmap confirms that holiday effects should not be interpreted as fixed constants. Instead, they should be treated as contextual variables whose impact may depend on the surrounding calendar, the commercial environment, and potential overlap with promotions or seasonal demand.

4.1.3.4 Relationship with Top-selling Weeks

Another useful perspective is whether holidays coincide with the strongest sales weeks of the year. If holidays were consistently positive demand drivers, holiday weeks would be expected to overlap frequently with the annual top-selling weeks. Nevertheless, the overlap analysis suggests that this is rarely the case.

There is no overlap in 2022, 2024, and 2025, while in 2023 only two holiday weeks appear among the top-selling weeks of the year: Clean Monday and Ochi Day. This indicates that holidays do not generally explain the highest demand peaks in the dataset. Instead, they function primarily as contextual calendar variables and, in many cases, as short-run suppressors of weekly sales.

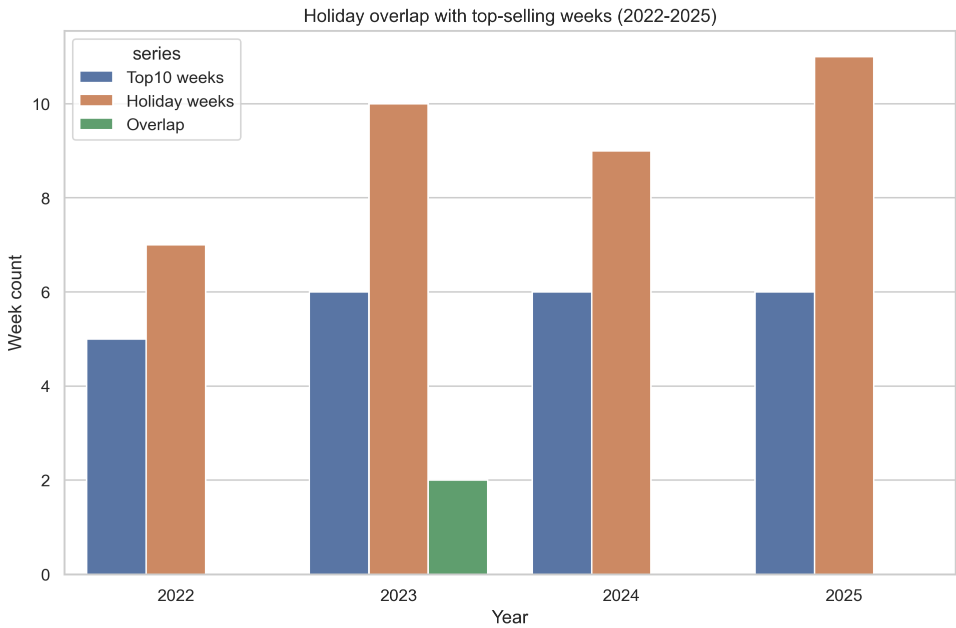

Figure 4.10: Overlap between holiday weeks and annual top-selling weeks

Taken together, the holiday results imply that calendar effects should be incorporated into the forecasting framework in a selective and event-specific manner. A single binary variable indicating the presence of any holiday would be too broad, because it would combine holidays with positive, neutral, and strongly negative effects into one category. A more appropriate approach is to use holiday-specific indicators or small holiday groups with similar empirical behaviour. For example, Clean Monday and Ochi Day may be treated separately because they exhibit relatively stronger upside potential, whereas Dormition, New Year, Boxing Day, and Easter-related holidays may be modelled as negative or disruption-related calendar controls.

4.1.4 Lag Features

Lag features are included to capture the recent history of demand before each forecast origin. In time-series forecasting, past values of the target variable often contain useful information about the current and future level of demand. This is especially relevant in weekly SKU-level forecasting, where recent sales movements may reflect changes in customer behaviour, product availability, or short-term demand momentum.

The main lag features considered in this analysis are previous weekly sales values, such as one-week, two-week, four-week, and longer seasonal lags. These variables allow the model to observe recent demand levels without using future information. Also, rolling statistics such as moving averages can be used to summarise recent demand over a fixed historical window, for example, a four-week rolling mean may capture short-term demand momentum, while a longer rolling window may provide a smoother estimate of the baseline demand level.

Lag features must be constructed carefully in order to avoid data leakage. Only information that would have been available at the time of forecasting can be used . So, when rolling-origin backtesting is applied, lag and rolling features should be recalculated at each forecast origin using only past observations ensuring that the evaluation remains realistic and that the model does not indirectly use future sales information.

Lag features are mainly relevant for regression-based or promo-aware extensions of the forecasting framework. Classical models such as ETS and SARIMA already use past demand through their internal time-series structure, while Croston-type methods use past non-zero demand sizes and intervals. Nevertheless, explicit lag features remain useful when external drivers such as promotions and holidays are combined with historical demand information in a regression-style model.

4.1.5 Price Indexing Features

Price-related information is another potential feature group because changes in price or discount intensity may influence demand. Nonetheless, price data are often less complete or less directly observable than sales quantities and promotion indicators and for this reason, price indexing features are treated cautiously in this thesis and are considered only where turnover and quantity information allow a consistent price proxy to be constructed.

A simple SKU-level unit price proxy can be calculated by dividing sales value by sales quantity for a given period. This unit price can then be indexed relative to a baseline level, such as the SKU's average price or the category-level average price. The resulting price index does not necessarily represent the exact shelf price paid by consumers, but it can provide a useful approximation of relative price movement over time.

Price indexing may be particularly useful when combined with promotion information. For example, a promotion flag identifies whether a campaign was active, while a price index may help indicate whether the campaign was associated with a lower effective price. This distinction matters because not all promotions operate in the same way. A `1+1` campaign, a price-off action, and a promotion stand may create different demand responses even if they are all recorded as promotional events.

Taking that into consideration, price features should only be included in the final forecasting specification if they are reliable, consistently available, and known before the forecast is made. If the price proxy is unstable or derived from the same realised sales values that the model is trying to predict, it may introduce noise or leakage and it has to be treated as a candidate feature rather than an automatic input for all models.

4.1.6 ABC–XYZ Segmentation and SKU Classification

ABC–XYZ segmentation is used to classify SKUs according to both business importance and demand behaviour as mentioned before as well. This is important because the product portfolio is heterogeneous: some SKUs contribute strongly to total sales value and require more attention, while others have low volume, irregular demand, or frequent zero-sales weeks. As discussed in the literature review and in the data part, segmentation helps connect forecasting choices with operational priorities and prevents the use of one universal method across all products (Flores and Whybark 1986; Syntetos, Boylan, and Croston 2005).

ABC classification ranks SKUs according to value contribution.The segmentation procedure is mentioned before so for example, A-items may require more careful monitoring because forecasting errors for these products have larger operational and financial consequences.

XYZ classification captures demand variability and stability as previously stated as well, this classification is especially important because many weekly SKU series contain zero-sales periods. Combining XYZ classification with arrival rate and demand variability helps identify which SKUs may be suitable for classical time-series methods and which may require intermittent-demand methods such as Croston or SBA.

The combination of ABC and XYZ creates a practical segmentation framework. For example, an AX item is both important and relatively stable, so it may be suitable for methods such as ETS or SARIMA. An AZ item is important but irregular, meaning that forecasting may be more difficult and that special attention should be paid to zero-sales weeks, promotional effects, or aggregation. Lower-priority CZ items may not justify complex modelling and may be better handled through simple rules or intermittent-demand approaches.

In the final forecasting framework, ABC–XYZ segmentation is used mainly for interpretation, model assignment, and segment-level reporting. It allows forecast performance to be evaluated not only at the total portfolio level but also by SKU type. This is useful because a model that performs well overall may still perform poorly for specific segments, especially for sparse or highly volatile products.

4.1.7 Final Feature Selection

The final feature selection is based on three criteria: operational relevance, data reliability, and forecasting usefulness. A feature is considered useful only if it can be known before the forecast is made, can be constructed consistently across the dataset, and has a meaningful relationship with short-term demand. This prevents the model from becoming unnecessarily complex and ensures that the forecasting framework remains transparent and usable.

The first group of selected features consists of historical demand information, including lagged sales and rolling demand summaries where appropriate. These variables capture the internal time-series structure of each SKU and provide the foundation for short-term forecasting. The second group consists of promotion variables, including promotion flags, promotion type indicators, duration, frequency, and recency measures. These features are retained because the descriptive analysis shows that promotions are frequent, heterogeneous, and associated with short-term changes in demand.

The third group consists of holiday and calendar indicators. However, the holiday analysis shows that calendar effects are not uniform. Therefore, holiday variables should be included selectively, either through holiday-specific indicators or through grouped indicators for holidays with similar empirical behaviour. This avoids treating all holidays as identical and allows the model to distinguish between potentially positive, neutral, and negative calendar effects.

The fourth group consists of price indexing features, but these are treated as conditional features. They should be included only if the constructed price proxy is reliable and does not create leakage. If price information is incomplete or unstable, price features should remain part of the exploratory analysis rather than the core model specification.

Finally, ABC–XYZ segmentation is retained as a structural feature of the forecasting framework. It is not only a modelling input but also a way to organise results, assign candidate models, and interpret performance across different SKU groups. The final feature set therefore combines historical demand, selected event variables, and segmentation logic in order to support a forecasting framework that is both data-driven and operationally interpretable.

4.2 Dataset merging and final structure

After mentioning all the features and data preprocessing it is now time to have a clear point of view of the final dataset and structure. Following construction of the weekly sales panel, the final analytical dataset is compiled by merging the weekly sales table with the promotion feature table and the product segmentation output. The merge is performed at `SKU-week` level by using the product identifier (`SKU`) and the weekly reference date (`weekstart`) as the main matching keys. This ensured that weekly sales represented the unit of analysis, and explanatory information is added consistently to each record. More specifically, time index, product descriptors and weekly sales values are contributed by the weekly sales table and engineered features such as promotion presence, campaign duration, the maximum and average promotional uplift, promotion peak turnover, campaign start and end dates, promotion-type indicators and recency-related variables, such as the previous promotion week and number of weeks since the previous promotion, are added in the promotion table. Furthermore, holiday indicators and product segment labels are also appended to the weekly panel so the subsequent `SKU-week` records capture sales, promotion conditions, calendar effects, and the overall behavioural category for the product. The resulting merged file has a panel structure in which an observation corresponds to one product in one specific week. Whenever a promotion goes on for longer than one week, the campaign information is replicated across all weekly rows within the campaign period; weekly sales vary from one week to the next. This architecture makes the dataset amenable to supervised forecasting, preserving the chronological structure of the dependent variable while enriching each observation with covariates which in turn line up to show how the dependent variable behaves on that basis.Below stands a small preview of two products in otder to have an exact preview of the merged dataset.

Table 4.4: Preview of the merged weekly sales, promotion, holiday, and segmentation dataset

SKU	Week start	Year	Week	Product	Sales	Holiday	Promo	Dur.	Segment
22000141	2022-06-13	2022	24	(J) TBR EXPERT WHITE (SENS)	47	Holy Spirit Monday	1	2	BZ
22000141	2022-06-20	2022	25	(J) TBR EXPERT WHITE (SENS)	75		1	2	BZ

Chapter 5

Methodology and Evaluation Framework

5.1 Methodology and Evaluation Framework

This chapter presents the methodology used to generate and evaluate weekly SKU-level forecasts for Just Health. The forecasting problem involves multiple time series with different demand patterns, so the framework is designed to compare candidate methods and assign them according to SKU segment rather than applying a single model to the entire portfolio.

The chapter also combines benchmark forecasting methods, intermittent-demand methods, and event-based adjustments within a segment-aware evaluation setup. Forecast performance is assessed using a fixed 12-week holdout horizon and measured with MAE and WAPE at SKU level before being aggregated by segment. The following sections explain the methodology of each baseline model and the evaluation metrics and procedure in more detail.

5.1.1 Evaluation Metrics

Forecast accuracy is evaluated using Mean Absolute Error and Weighted Absolute Percentage Error. These metrics provide complementary views of forecast performance. MAE measures the average absolute error in the same unit as the original demand series. WAPE expresses the total forecast error relative to total actual demand and is therefore useful for segment-level comparison.

MAE is defined as:

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|,$$

where y_t is the observed demand, $\hat{y}_t$ is the forecast, and n is the number of forecasted observations.

WAPE is defined as:

$$WAPE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{\sum_{t=1}^{n} |y_t|}.$$

When reported as a percentage, WAPE is multiplied by 100:

$$WAPE(\%) = WAPE \times 100.$$

MAE is useful because it is easy to interpret in demand units and it is also scale-dependent, meaning that it cannot always be compared directly across SKUs with different demand volumes. WAPE is used as the main comparison metric because it expresses error relative to the total realised demand. Thus, its more suitable for comparing model performance across segments.

Special care is needed when interpreting WAPE for highly intermittent or very low-volume segments. If total realised demand is very small, even small absolute errors can generate very high WAPE values. This explains why some sparse segments, especially CZ, produce extreme WAPE values across several models. For this reason, the evaluation considers both numerical accuracy and the demand characteristics of each segment.

5.1.2 Forecasting Framework

The forecasting framework is designed to generate weekly SKU-level demand forecasts for operational planning. The process starts from the cleaned weekly sales dataset, where each observation corresponds to one SKU in one calendar week. The forecasting target is the weekly sales variable denoted as `sales_raw_weekly`. Each SKU series is ordered chronologically and then divided into a training period and a fixed test horizon.

Other than historical sales, the dataset contains engineered variables related to promotions and holidays. These variables are included because, as discussed in the literature review and feature analysis chapter, promotions and calendar effects can create temporary demand deviations that are not fully explained by the internal time-series pattern alone and they can genearate better results. In the final implementation, these event variables are used in two different ways. For SARIMA, they are included directly as exogenous regressors. For Croston and SBA, they are applied through a post-forecast uplift adjustment.

The general forecasting problem can be expressed as follows. For each SKU i and week t, the objective is to forecast future weekly demand $y_{i,t+h}$ for horizons $h = 1, \ldots, 12$, using historical demand observations and, where applicable, known event variables:

$$\hat{y}_{i,t+h} = f(y_{i,1}, y_{i,2}, \ldots, y_{i,t}, X_{i,t+h}),$$

where $\hat{y}_{i,t+h}$ is the forecast for SKU i at horizon h, and $X_{i,t+h}$ represents available external features such as promotion and holiday indicators.

The final forecasting framework is operationalised as a segment-based routing system. This means that the final forecast is not generated by one universal model for all SKUs, but by a structured decision rule that maps each segment to the model that performed best or most appropriately for that segment. This design balances interpretability, robustness, and practical usability, since the model choice can be explained in relation to the demand behaviour of each SKU group.As we will see below, based on the results of each baseline model the segmentation will occur in order to achieve the best possible results.

For non-SARIMA models, the framework also applies a promotion and holiday adjustment layer. First, the base forecast is generated using the selected intermittent-demand method. Then, event uplift factors are estimated from the training data and applied to the forecast horizon when promotions or holidays are present. This allows the framework to preserve the simplicity of Croston and SBA while still incorporating short-run commercial and calendar effects.

5.1.3 Baseline Forecasting Models

The baseline model comparison includes five forecasting methods: Seasonal Naïve, ETS, Croston, SBA, and SARIMA. These methods were selected because they represent different levels of complexity and are suitable for different types of demand behaviour. Seasonal Naïve provides a simple seasonal benchmark, ETS captures level, trend, and seasonality through exponential smoothing, SARIMA models seasonal autocorrelation, and Croston/SBA address intermittent demand.

All models are evaluated under the same general forecasting objective and compared using the same accuracy metrics. This ensures that performance differences are due to the forecasting methods themselves rather than to different evaluation conditions. The following subsections describe each method and present its segment-level performance.

5.1.3.1 Seasonal Naïve Method

The Seasonal Naïve method serves as the fundamental benchmark in this analysis. It assumes that the value of a future observation will be equal to the value observed in the corresponding period of the previous seasonal cycle and since the data are weekly, the seasonal period is set to 52 weeks.

The Seasonal Naïve forecast is defined as:

$$\hat{y}_{t+h} = y_{t+h-s},$$

where $s = 52$ for weekly yearly seasonality, and h is the forecast horizon.

This method is intentionally simple. Its role is to establish a baseline performance level against which more advanced models can be evaluated. If a more complex model does not improve on Seasonal Naïve, then its additional complexity is difficult to justify.

5.1.3.2 ETS Method

The Exponential Smoothing method is included as a classical benchmark for smoother time series. As mentioned before, ETS models are designed to capture level, trend, and seasonal components by applying exponentially decreasing weights to past observations. This allows the model to respond more strongly to recent demand patterns while still retaining information from the historical series.

A simple exponential smoothing formulation can be written as:

$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1},$$

where ℓ_t is the smoothed level at time t, y_t is the observed demand, and α is the smoothing parameter. More general ETS specifications can also include trend and seasonal components.

5.1.3.3 Croston Method

The Croston method is implemented to address the limitations of standard time-series methods under intermittent demand. Unlike models that smooth demand over all periods, Croston separates the estimation of demand size from the estimation of the time interval between non-zero demand observations. This makes it more suitable for SKUs with many zero-sales weeks.

Let z_j denote the size of the j-th non-zero demand observation and p_j the interval between two non-zero demand observations. Croston's method updates these two components separately:

$$\hat{z}_j = \alpha z_j + (1-\alpha)\hat{z}_{j-1},$$

$$\hat{p}_j = \alpha p_j + (1-\alpha)\hat{p}_{j-1}.$$

The forecast is then calculated as:

$$\hat{y}_t = \frac{\hat{z}_j}{\hat{p}_j}.$$

5.1.3.4 SBA Method

The Syntetos–Boylan Approximation is a bias-corrected refinement of Croston's method. The motivation for using SBA is that the original Croston estimator is known to produce biased forecasts under certain intermittent-demand conditions. SBA applies a correction factor to reduce this bias while preserving the main logic of estimating demand size and demand interval separately.

In the implementation used in this thesis, SBA is calculated as a bias-corrected approximation based on the mean non-zero demand and the average interval between non-zero demand observations. Let $\hat{z}$ denote the average non-zero demand and $\hat{p}$ the average demand interval:

$$\hat{z} = \frac{1}{N^+}\sum_{t:y_t>0} y_t,$$

$$\hat{p} = \frac{T}{N^+},$$

where T is the number of observed training weeks and N^+ is the number of weeks with positive demand. The SBA forecast is then:

$$\hat{y}^{SBA} = \left(1 - \frac{\alpha}{2}\right)\frac{\hat{z}}{\hat{p}}.$$

Moreover, $\alpha = 0.1$ is used. The resulting forecast is repeated across the 12-week horizon and restricted to non-negative values.

5.1.3.5 SARIMA Method

The Seasonal Autoregressive Integrated Moving Average model is included to capture seasonal autocorrelation and recurring patterns in the weekly demand series. SARIMA extends the ARIMA framework by including seasonal autoregressive, differencing, and moving-average components. This makes it suitable for SKUs with sufficient historical depth and stable seasonal structure.The main idea is that SARIMA will have the best results out of all the other baseline models but this will not be so easy because of the small amount of data that is in the company's database.

A general SARIMA model can be expressed as:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t,$$

where B is the backshift operator, s is the seasonal period, ϕ_p and θ_q are the non-seasonal autoregressive and moving-average polynomials, and Φ_P and Θ_Q are the seasonal autoregressive and moving-average polynomials.

When exogenous variables are included, the model can be expressed as:

$$y_t = \beta X_t + n_t,$$

where X_t contains external regressors such as promotion and holiday variables, and n_t follows a SARIMA process.

In the final implementation, SARIMA is applied only to segment AY and only when the SKU has enough training history. The seasonal period is set to 52 weeks. A limited set of candidate SARIMA specifications is tested and the specification with the lowest AIC is selected. This lightweight model-selection strategy keeps the procedure computationally manageable while still allowing some flexibility in the SARIMA structure.

5.1.4 Train-Test Split and Forecast Horizon

The evaluation uses a fixed holdout design. For every SKU, the last 12 weeks of available observations are held out as the test set, while all previous observations are used for training. This design reflects the practical objective of the thesis, which is to generate short-term forecasts for approximately three months ahead at weekly frequency.

For a time series of length T, the training set is defined as:

$$\{y_1, y_2, \ldots, y_{T-12}\},$$

while the test set is defined as:

$$\{y_{T-11}, y_{T-10}, \ldots, y_T\}.$$

The forecast horizon is therefore:

$$h = 1, 2, \ldots, 12.$$

This setup ensures that every model is evaluated under the same conditions. It also avoids evaluating models only on in-sample fit, which would overstate forecasting performance. By using a future holdout period, the evaluation better approximates the real operational situation in which forecasts are made before actual demand is observed.

Minimum history requirements are applied before fitting the models. A SKU must have at least 38 total weeks to be included in the forecasting loop. For SARIMA, at least 78 training weeks are required because seasonal and autocorrelation parameters cannot be estimated reliably from very short series. For SBA, at least 26 training weeks and at least 4 non-zero training weeks are required. For Croston, at least 20 training weeks and at least 2 non-zero training weeks are required. If these requirements are not satisfied, the model uses the last-value naïve fallback forecast.

5.1.5 Implementing Promo Features and Holiday Effects

The dataset encodes promotional activity through a combination of binary indicators and continuous variables. A general promotion flag identifies whether any promotional activity was active during a given week, while additional variables capture the number of concurrent promotions, their duration in weeks, and an overall intensity score reflecting the magnitude of the promotional effect relative to normal trading conditions. Further type-specific binary flags distinguish between different promotional mechanics, such as leaflet campaigns, bundle discounts, multi-buy offers, and in-store display activities. A separate lag indicator marks the week immediately following a promotion, capturing potential post-promotional demand dip effects. Holiday information is represented as a binary weekly flag indicating whether a Greek public holiday falls within that week, accompanied by a descriptive text label used solely for interpretability and not as a modelling input.

5.1.6 Differentiated Treatment Across Models

The way these features are incorporated into the forecasting pipeline differs depending on the model assigned to each segment. For SARIMA, all promotional variables and the holiday indicator are passed directly as exogenous regressors during both the training and the forecasting steps. The model is therefore estimated as:

$$\Phi(B)\tilde{\Phi}(B^s)\,\Delta^d\Delta_s^D\,y_t \;=\; \boldsymbol{x}_t^\top\boldsymbol{\beta} \;+\; \Theta(B)\tilde{\Theta}(B^s)\,\varepsilon_t \tag{5.1}$$

where $\boldsymbol{x}_t$ is the vector of exogenous inputs at week t (promotional and holiday indicators), $\boldsymbol{\beta}$ is the corresponding coefficient vector estimated during training, B is the backshift operator, $s = 52$ is the seasonal period, and ε_t is white noise. This formulation allows SARIMA to learn the direct statistical relationship between each feature and sales demand, naturally incorporating their effect into the 12-week forecast provided that the future values of these regressors are available in the test window.

Uplift-Based Adjustment for SBA and Croston

For SBA and Croston, neither model natively supports external regressors, as both methods operate exclusively on the demand series. This was really challenging at first but to overcome this limitation, a post-hoc multiplicative uplift adjustment is applied after the base forecast is generated. Let μ_0 denote the baseline mean demand, defined as the average sales during non-promotional, non-holiday weeks in the training period. The three uplift factors are then estimated as:

$$\delta_{\text{promo}} \;=\; \frac{\mu_{\text{promo}}}{\mu_0}, \qquad \delta_{\text{holiday}} \;=\; \frac{\mu_{\text{holiday}}}{\mu_0}, \qquad \delta_{\text{both}} \;=\; \frac{\mu_{\text{promo}\cap\text{holiday}}}{\mu_0} \tag{5.2}$$

where μ_{promo}, μ_{holiday}, and $\mu_{\text{promo}\cap\text{holiday}}$ are the mean sales observed during promotional weeks, holiday weeks, and weeks where both events coincide, respectively. To prevent extreme adjustments driven by outlier observations, each factor is clipped to a bounded range:

$$\delta_{\text{promo}} \in [0.8,\ 2.0], \qquad \delta_{\text{holiday}} \in [0.8,\ 1.5], \qquad \delta_{\text{both}} \in [0.8,\ 2.2] \tag{5.3}$$

Week-Level Adjustment Logic

Once the uplift factors are estimated, the adjusted forecast for each week t in the horizon is computed as:

$$\hat{y}_t^* \;=\; \hat{y}_t \;\times\; \begin{cases} \delta_{\text{both}}, & \text{if promotion and holiday are both active and } \delta_{\text{both}} \text{ exists} \\ \delta_{\text{promo}} \cdot \delta_{\text{holiday}}, & \text{if both flags active but } \delta_{\text{both}} \text{ unavailable} \\ \delta_{\text{promo}}, & \text{if only promotion is active} \\ \delta_{\text{holiday}}, & \text{if only holiday is active} \\ 1, & \text{otherwise} \end{cases} \tag{5.4}$$

where $\hat{y}_t$ is the base forecast produced by SBA or Croston. An additional intensity-based correction is applied for active promotional weeks, scaling the forecast by a factor:

$$\gamma_t \;=\; 1 \;+\; \min\!\left(\frac{\phi_t}{1000},\; 0.20\right) \tag{5.5}$$

where ϕ_t is the promotional intensity score at week t. This caps the additional intensity-driven uplift at 20%, preventing over-correction in unusually high-intensity promotional weeks. The final adjusted forecast is then clipped to non-negative values:

$$\hat{y}_t^{\text{final}} \;=\; \max(\hat{y}_t^* \cdot \gamma_t,\; 0) \tag{5.6}$$

5.1.7 Main Forecasting Pipeline

Algorithm 1 Segment-based forecasting and evaluation pipeline

Load the weekly SKU-level forecasting dataset.

Standardise column names and convert date, target, promotion, and holiday variables to the correct data types.

Keep only valid segments $\{AY, AZ, BY, BZ, CY, CX\}$ and exclude segment CZ.

for each SKU i in the dataset **do**

  Sort the SKU observations chronologically by week.

  if the SKU has fewer than the minimum required total weeks **then**

    Exclude the SKU from the forecasting loop.

    continue

  end if

  Identify the SKU segment g_i.

  Assign the forecasting model according to the segment mapping:

$$m(i) = \begin{cases} SARIMA, & \text{if } g_i = AY, \\ SBA, & \text{if } g_i \in \{AZ, BY, BZ, CY\}, \\ Croston, & \text{if } g_i = CX \\ ETS, & \text{if } g_i = CZ \end{cases}$$

  Split the SKU series into training data and a 12-week test horizon:

$$Train_i = \{y_{i,1}, \ldots, y_{i,T-12}\}, \quad Test_i = \{y_{i,T-11}, \ldots, y_{i,T}\}.$$

  Count the number of non-zero demand weeks in the training set.

Algorithm 2 Segment-based forecasting and evaluation pipeline (continued)

if $m(i) = SARIMA$ **then**
  if the training history is shorter than the SARIMA minimum requirement **then**
    Generate a last-value naïve fallback forecast.
  else
    Fit candidate SARIMA specifications with promotion and holiday regressors.
    Select the SARIMA specification with the lowest AIC.
    Generate a 12-week SARIMA forecast.
  end if
else if $m(i) = SBA$ **then**
  if the training history or number of non-zero weeks is insufficient **then**
    Generate a last-value naïve fallback forecast.
  else
    Generate a 12-week SBA forecast.
  end if
  Apply promotion and holiday uplift adjustment.
else if $m(i) = Croston$ **then**
  if the training history or number of non-zero weeks is insufficient **then**
    Generate a last-value naïve fallback forecast.
  else
    Generate a 12-week Croston forecast.
  end if
  Apply promotion and holiday uplift adjustment.
else if $m(i) = ETS$ **then**
  if the training history or number of non-zero weeks is insufficient **then**
    Generate a last-value naïve fallback forecast.
  else
    Generate a 12-week ETS forecast.
  end if
  Apply promotion and holiday uplift adjustment.
else
  Generate a last-value naïve fallback forecast.
end if
Restrict forecasts to non-negative values.
Store actual demand, base forecast, final forecast, forecast errors, model used, and fallback status.
Compute SKU-level MAE and WAPE.
end for
Aggregate SKU-level metrics by segment and model.
Export forecast details, SKU metrics, segment metrics, excluded SKUs, and run summary files.

In order to have a clear explanation to the above the algorithm begins by loading the weekly SKU-level dataset. Then it cleans and standardises the columns (names and data types) and keeps only the relevant segments (AY, AZ, BY, BZ, CY, CX), excluding CZ due to its unreliable demand pattern and that happened because the results where not the expected and was very difficult to lower the evaluation metrics.

Regarding the iteration over each SKU for each product individually, the observations are sorted chronologically. If a SKU has too few weeks of data, it is excluded entirely and the algorithm moves on to the next one. This happened because the best results need to be objective so it is needed enough data. It will then become apparent that, for this reason, some segments are missing from the final results

Then on model assignment based on segment each SKU belongs to a segment, and based on that segment the appropriate forecasting model is automatically assigned based on the results of baseline models that are presenting on the result section that define which model is best for every segment:

SARIMA for AY (stable, seasonal demand)

SBA for AZ, BY, BZ, CY (intermittent demand)

Croston for CX (also intermittent but of a different type)

ETS for CZ (but it is excluded due to demand pattern)

Afterwards there is Train/test split, means for each SKU's data is split into a training set (the full history excluding the last 12 weeks) and a test set (the final 12 weeks), against which the forecast is evaluated.

On model fitting and fallback logic for each model, the algorithm checks whether there is sufficient data to train it reliably:

If yes → the model is trained normally and produces a 12-week forecast

If no → a simple fallback is used (the last known demand value)

For SARIMA specifically, multiple specifications are tested and the one with the lowest AIC is selected. For SBA and Croston, an additional uplift adjustment is applied to account for promotions and public holidays.

Furthermore on storing results, forecasts are clipped to non-negative values and stored alongside the actual demand, forecast errors, and a flag indicating whether a fallback was used. MAE and WAPE are computed for each SKU.

Finally, results are aggregated by segment and exported to output files: forecast details, SKU-level and segment-level metrics, excluded SKUs, and a run summary report.

Chapter 6

Results

6.1 Results

This chapter presents the baseline and final forecasting results obtained from the segment-based forecasting framework. The evaluation is based on SKU-level forecasts over a fixed 12-week test period as mentioned at the methodology chapter as well. The results are first presented at model level, then by SKU segment, followed by a comparison with the baseline model-selection stage and an interpretation of the main findings. The purpose of this chapter is to report accuracy values and to assess how forecasting performance differs across demand segments.

6.1.1 Baseline Results

6.1.1.1 Seasonal Naïve

The performance metrics, presented in Table 6.1 and Figure 6.1, show clear variation across product segments. Segments such as CX have very low error, while BZ and CZ exhibit very high WAPE values. This pattern is mainly driven by the intermittent nature of demand in these categories.

Table 6.1: Seasonal Naive Performance Metrics per Segment

Segment	**MAE**	**WAPE (%)**
AY	60.11	53.36
AZ	77.05	102.07
BY	57.01	51.89
BZ	23.84	231.02
CX	1.82	2.69
CY	40.89	57.60
CZ	14.26	296.06

The high WAPE values for BZ and CZ confirm that Seasonal Naïve struggles when demand is sparse. These results mean that sparse segments require more appropriate forecasting methods, such as Croston or SBA. The CZ segment, however, remains problematic across several methods and is therefore excluded from the final forecasting pipeline. This is not only due to the small volume of data, but also to the many zero sales figures and the highly volatile sales—which are either very low or very high.

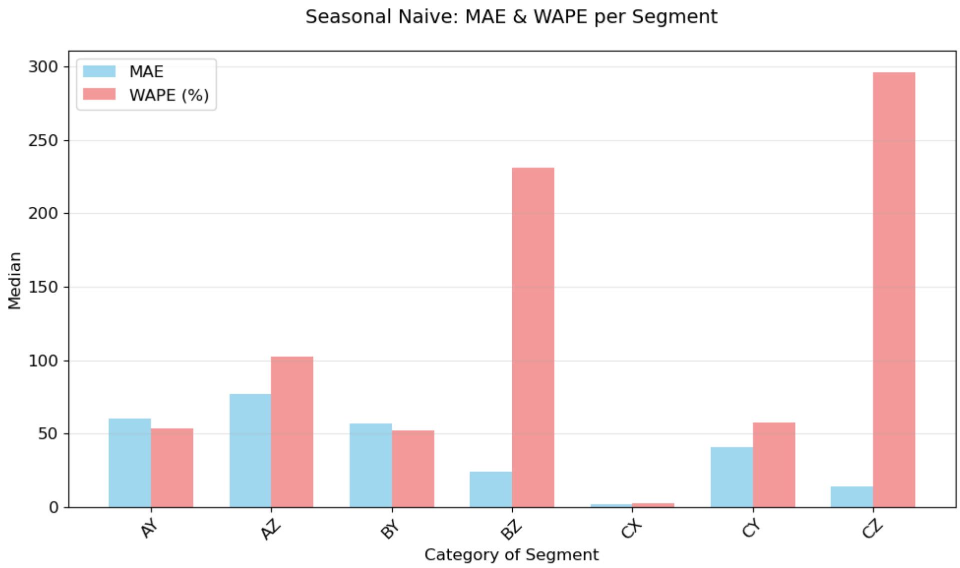

Figure 6.1: Mean MAE and WAPE distribution across product segments using the Seasonal Naive model.

6.1.1.2 ETS

A comparison of ETS results against Seasonal Naïve is provided in Table 6.2 and Figure 6.2. ETS improves performance in some segments, particularly where the series have smoother structure. However, the model still struggles in sparse segments such as BZ and CZ making it difficult to estimate stable smoothing components.

Table 6.2: ETS Performance Metrics per Segment

Segment	MAE	WAPE (%)
AY	40.41	34.96
AZ	58.18	70.86
BY	191.03	80.69
BZ	17.27	238.49
CX	1.22	1.84
CY	54.62	74.52
CZ	23.05	236.05

The results of ETS can provide useful adaptivity for some demand patterns, but it does not fully resolve the challenges created by intermittent demand. Therefore, ETS is retained as a relevant benchmark, but it is not selected as the final method for the evaluated segments due to the results it got which are not sufficient.

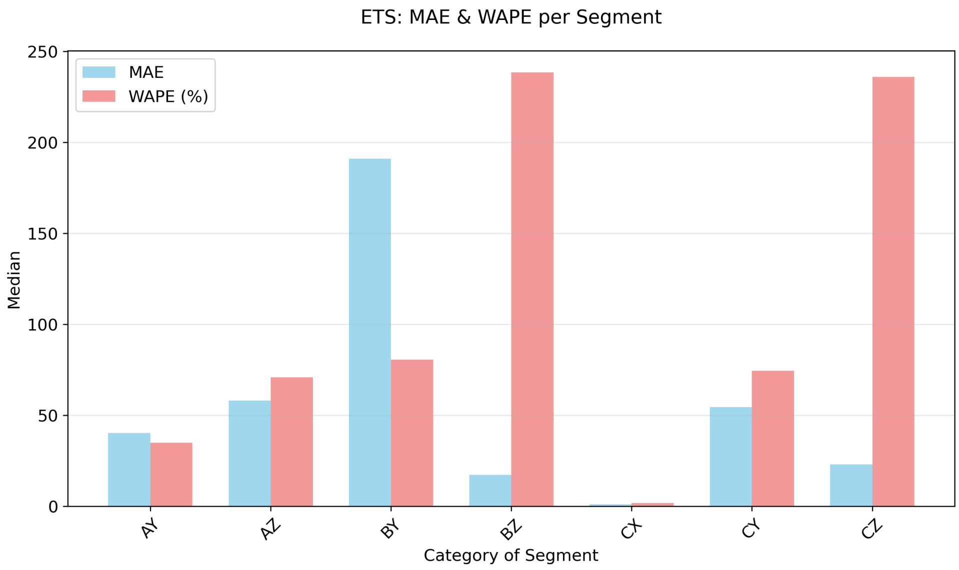

Figure 6.2: Mean MAE and WAPE distribution across product segments using the ETS model.

6.1.1.3 Croston

Moving to intermittent-demand methods, Table 6.3 and Figure 6.3 explain that Croston provides more balanced performance across several segments compared with Seasonal Naïve and ETS. In particular, the BZ segment improves substantially, which confirms that intermittent-demand methods are more appropriate for sparse series.

Table 6.3: Croston Performance Metrics per Segment

Segment	MAE	WAPE (%)
AY	38.07	34.40
AZ	66.71	65.20
BY	41.45	32.69
BZ	25.88	73.15
CX	1.07	1.61
CY	19.20	26.23
CZ	15.61	508.51

Croston is therefore an important candidate for intermittent segments. However, the CZ segment continues to show unstable results, suggesting that its demand pattern may be too sparse or irregular for reliable model-based forecasting within the available historical period.

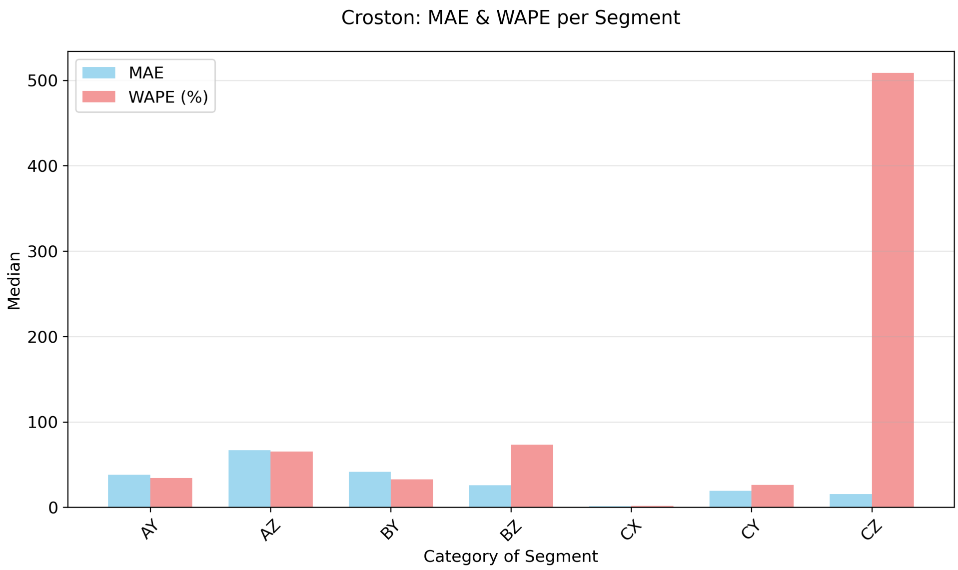

Figure 6.3: Mean MAE and WAPE distribution across product segments using the Croston model.

6.1.1.4 SBA

Building on the Croston framework, the SBA results are presented in Table 6.4 and Figure 6.4. The model provides strong performance across several segments, particularly AZ, BY, BZ, and CY , meaning that the bias-corrected intermittent-demand approach is well suited to the demand patterns observed in the dataset.

Table 6.4: SBA Performance Metrics per Segment

Segment	MAE	WAPE (%)
AY	30.70	27.72
AZ	53.09	51.78
BY	33.73	26.37
BZ	20.67	58.33
CX	1.42	3.36
CY	15.76	21.50
CZ	12.22	395.13

The improved stability of SBA in several segments supports its use as the preferred method for many SKU groups. In particular, SBA performs better than Seasonal Naïve and ETS in intermittent or moderately irregular segments, making it a strong candidate for the final segment-level model selection.

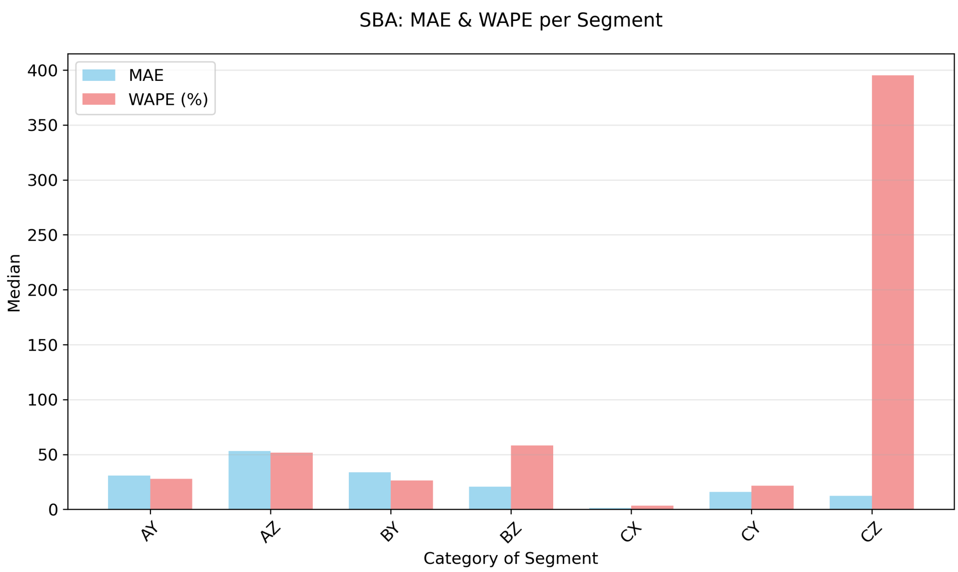

Figure 6.4: Mean MAE and WAPE distribution across product segments using the SBA model.

6.1.1.5 SARIMA

In contrast to the intermittent-demand methods examined above, Table 6.5 and Figure 6.5 illustrate how SARIMA behaves across the same segment portfolio. SARIMA performs particularly well in the AY segment, where the demand structure is more stable and suitable for seasonal modelling. On the other hand, its performance deteriorates in sparse and intermittent segments such as BZ and CZ as the other models.

Table 6.5: SARIMA Performance Metrics per Segment

Segment	**MAE**	**WAPE (%)**
AY	16.02	22.18
AZ	261.81	327.27
BY	180.00	225.00
BZ	187.84	234.80
CX	40.22	50.27
CY	62.79	78.48
CZ	203.97	254.97

These results of SARIMA is highly useful for specific stable segments but is not appropriate as a universal model for the entire portfolio. This has as result that SARIMA is selected only where the demand pattern and historical depth support seasonal modelling.

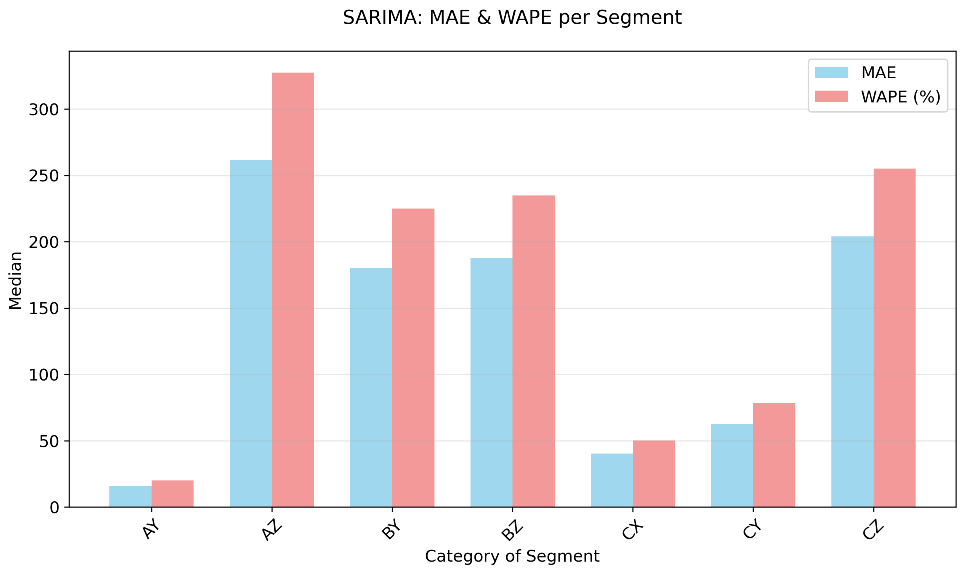

Figure 6.5: Mean MAE and WAPE distribution across product segments using the SARIMA model.

6.1.2 Model Comparison Procedure

The model comparison procedure combines the baseline model results into a final segment-level model selection. Each candidate model is evaluated for every valid segment using the same 12-week test horizon and the same accuracy metrics. The primary selection criterion is WAPE minimisation, while MAE is used as a complementary measure to interpret the absolute size of the errors.

Table 6.6 presents the top-performing model for each segment based on WAPE minimisation.

Table 6.6: Winning Forecasting Models per Segment

Segment	Winning Model	WAPE (%)
AY	SARIMA	22.18
AZ	SBA	51.78
BY	SBA	26.37
BZ	SBA	58.33
CY	SBA	21.50
CX	Croston	1.61
CZ	ETS	236.05

These results as expected mean that no single model dominates across all segments. SARIMA is selected for AY, where the demand structure is more stable and seasonally regular. SBA is selected for AZ, BY, BZ, and CY, indicating that the bias-corrected intermittent-demand approach performs well across several heterogeneous segments, but all segments will be used in the final results as mentioned. Croston is selected for CX, where it provides the lowest WAPE among the evaluated methods. The CZ segment is excluded from the final selected-model table because as previously stated has not have acceptable results.

The visual comparison of model performance is in Figure 6.6.

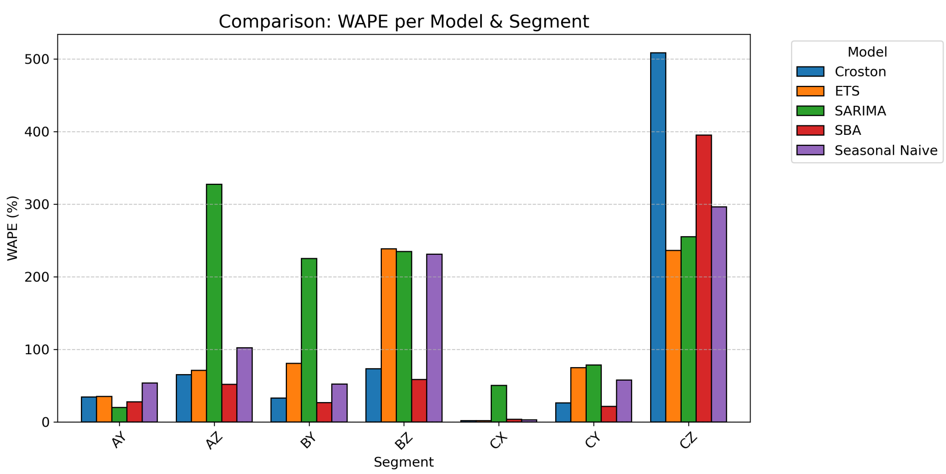

Figure 6.6: Comparative WAPE across all evaluated models per segment.

The final comparison confirms the need for a modular forecasting approach. Stable and seasonal segments benefit from models that capture temporal dependence, while intermittent segments require methods that account for irregular demand arrivals. This supports the final segment-based routing rule used in the implementation.

6.1.3 SKU-based Model Selection Approach

This format of the SKU model selection is necessary because the product portfolio contains items with very different demand behaviours. Some SKUs exhibit stable and seasonal patterns, while others are sparse, intermittent, or dominated by zero-demand weeks.

The baseline comparison stage evaluates several candidate models across the valid SKU segments. The final implementation then uses a fixed segment-to-model mapping based on the observed segment-level results and the demand characteristics of each group. In the final forecasting pipeline, the mapping rule is:

$$m(i) = \begin{cases} SARIMA, & \text{if } segment_i = AY, \\ SBA, & \text{if } segment_i \in \{AZ, BY, BZ, CY\}, \\ Croston, & \text{if } segment_i = CX \\ ETS, & \text{if } segment_i = CZ. \end{cases}$$

The segment-level model selection rule can be written as:

$$m_g^* = \arg\min_{m \in M} \text{WAPE}_{g,m},$$

where m_g^* is the selected model for segment g, M is the set of candidate forecasting models, and $\text{WAPE}_{g,m}$ is the weighted absolute percentage error of model m for segment g.

The final implementation also includes minimum-history requirements. These thresholds prevent models from being fitted on series that are too short or too sparse to support meaningful estimation. When a SKU does not satisfy the minimum requirements for its assigned model, the system falls back to a last-value naïve forecast:

$$\hat{y}_{t+h} = \max(y_t, 0), \quad h = 1, \ldots, 12.$$

This fallback mechanism ensures that the pipeline can still generate non-negative forecasts for short or unstable series, while clearly recording that a fallback was used.

6.1.4 Main Forecasting Results

The final forecasting results were evaluated by averaging the SKU-level metrics across all series assigned to each forecasting method. The final evaluation includes 172 SKUs and 2,064 forecast observations, corresponding to a 12-week test horizon for each SKU. Since each SKU belongs to a specific ABC–XYZ segment and each segment is routed to a selected forecasting method, the model-level results should be interpreted as the performance of the final segment-based forecasting framework rather than as a direct comparison between algorithms applied to the same identical set of series.

Table 6.7 presents the average results by forecasting model. The SBA-routed group was applied to the majority of the evaluated SKUs and achieved an average MAE of 10.45 and WAPE of 7.84%. The SARIMA-routed group achieved an average MAE of 28.63 and WAPE of 17.02%. These results of the final framework performs better, in relative terms, for the SKU groups routed to SBA than for the AY group routed to SARIMA. However, this difference should be interpreted carefully, because the two model groups correspond to different demand segments with different demand scales and demand regularity.

Table 6.7: Average forecasting metrics by model

Model	Train weeks	Nonzero train weeks	Zero ratio	MAE	WAPE (%)
SBA	145	84	0.442	10.45	7.84
SARIMA	159	144	0.090	28.63	17.02

The SARIMA-routed group has a longer training history, more non-zero training weeks, and a lower zero ratio. This means that the AY segment contains more continuous demand histories than the SBA-routed segments. Nevertheless, the SARIMA group also has a higher MAE, which suggests that the SKUs in this group operate at a larger sales scale. In contrast, the SBA-routed segments contain more zero-heavy demand patterns, but their lower demand scale leads to smaller absolute errors. This is why MAE and WAPE must be interpreted together.

The model-level results implement that the baseline demand structure remains the dominant driver of forecasting accuracy. Promotions and holidays provide useful short-term adjustments, but they do not fully determine the quality of the forecast. Their contribution is mainly corrective: they help adjust the baseline forecast around event weeks, but they cannot compensate for a weak or highly unstable historical demand signal. Therefore, the final forecasting accuracy depends on the combination of model choice, segment structure, demand scale, and the amount of usable historical information available for each SKU.

6.1.5 Note on the CX and CY Segments

It should be noted that the CX and CY segments, although they are included in the baseline model comparison and assigned to the Croston method based on its lowest WAPE of 1.61%, and SBA method as well with WAPE of 21.50 %, they do not appear in the final forecasting results. As discussed in the methodology chapter, the available data for CX and CY SKUs did not satisfy the minimum history requirements needed to support a reliable train-test split within the defined 12-week evaluation horizon. Specifically, the combination of a short overall observation window and a high proportion of zero-demand weeks left insufficient non-zero training observations for the Croston and SBA estimator to produce stable demand-interval estimates. As a result, CX and CY SKUs were either excluded from the final forecasting loop or routed to the fallback mechanism, and are therefore not represented in the segment-level summary table.Furthermore the results where incredibly accurate meaning that likely driven by insufficient data rather than true predictive performance.

6.1.6 Results by SKU Segment

The segment-level results provide a more detailed view of forecasting performance. This is necessary because model-level averages can hide important differences between SKU groups. In particular, the final corrected results contain four evaluated segments: AY, AZ, BY, and BZ. Table 6.8 reports the final performance by segment.

Table 6.8: Final forecasting metrics by SKU segment

Segment	Model	SKUs	Train weeks	Nonzero weeks	Zero ratio	MAE	WAPE (%)
AY	SARIMA	28	159	144	0.090	28.63	17.02
AZ	SBA	62	145	84	0.442	18.84	13.76
BY	SBA	1	145	84	0.442	3.55	1.96
BZ	SBA	81	145	84	0.442	4.12	3.38

The AY segment contains 28 SKUs and is the only segment routed to SARIMA in the final framework. It has the strongest demand continuity among the evaluated segments, with an average of 144 non-zero training weeks and a zero ratio of 0.090. However, it also records the highest MAE, which indicates that these SKUs are likely to have larger sales volumes and stronger weekly movements. The WAPE value of 17.02% is relative to realised demand, the forecasts remain within a reasonable range for short-term planning.

The AZ segment contains 62 SKUs and is forecasted using SBA. Its WAPE of 13.76% is lower than AY but higher than BY and BZ. This is because AZ items remain relatively difficult to forecast, even when an intermittent-demand method is used. This is expected because AZ products combine higher business importance with irregular or sparse demand behaviour.

The BY segment contains only one SKU in the final corrected results. Although its WAPE of 1.96% is very low, this result should be interpreted cautiously because it is based on a single product and therefore cannot support broad conclusions about the whole BY segment.

The BZ segment contains 81 SKUs and records a low average MAE of 4.12 and WAPE of 3.38%. This indicates that SBA performs well for this group in the final evaluation. Since the results for BZ are relatively strong, it is useful to note that they still correspond to zero-heavy products with low average demand volumes. In such cases, small absolute deviations can remain associated with low WAPE if the realised demand is also very small. The low WAPE is therefore encouraging, but it should be interpreted together with the segment's demand scale.

Figure 6.7 presents the mean MAE across the final evaluated SKU segments. The results of AY and AZ have clearly higher absolute forecast errors than BY and BZ. This pattern should be interpreted in relation to the scale of demand in each segment, because MAE is measured in the same units as the original sales variable and is therefore sensitive to SKU volume.

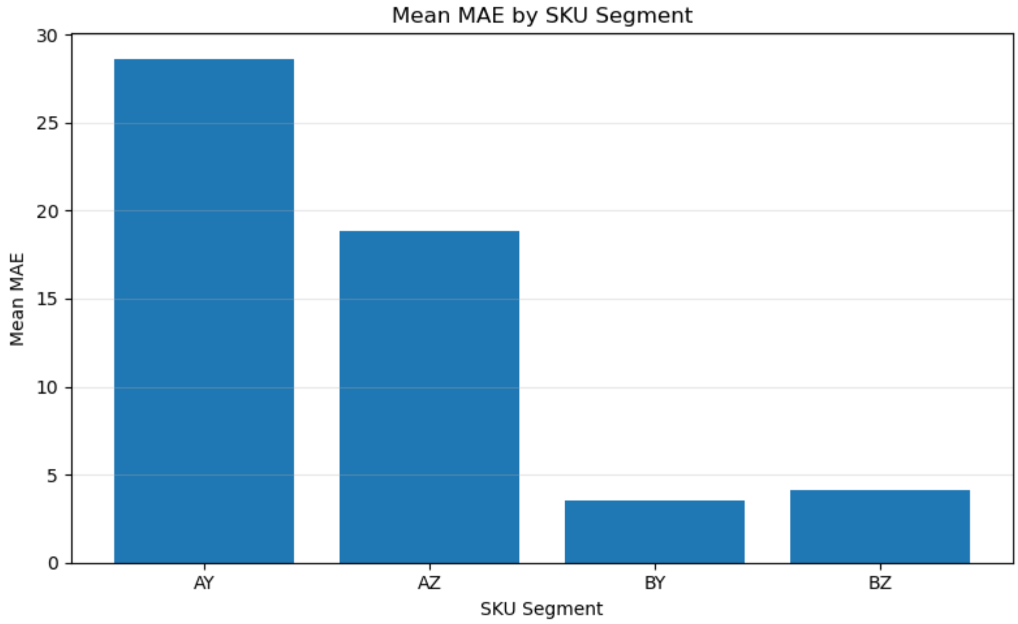

Figure 6.7: Mean MAE by SKU segment

Figure 6.8 complements the MAE results by showing the mean WAPE across SKU segments. Since WAPE expresses forecast error relative to realised demand, it provides a more comparable measure across segments with different sales scales. In order to explain the figure the AY and AZ have the highest relative errors again , while BY and BZ record substantially lower WAPE values. The BY result should again be interpreted cautiously because it is based on only one SKU. The comparison between Figures 6.7 and 6.8 highlights why both metrics are needed.

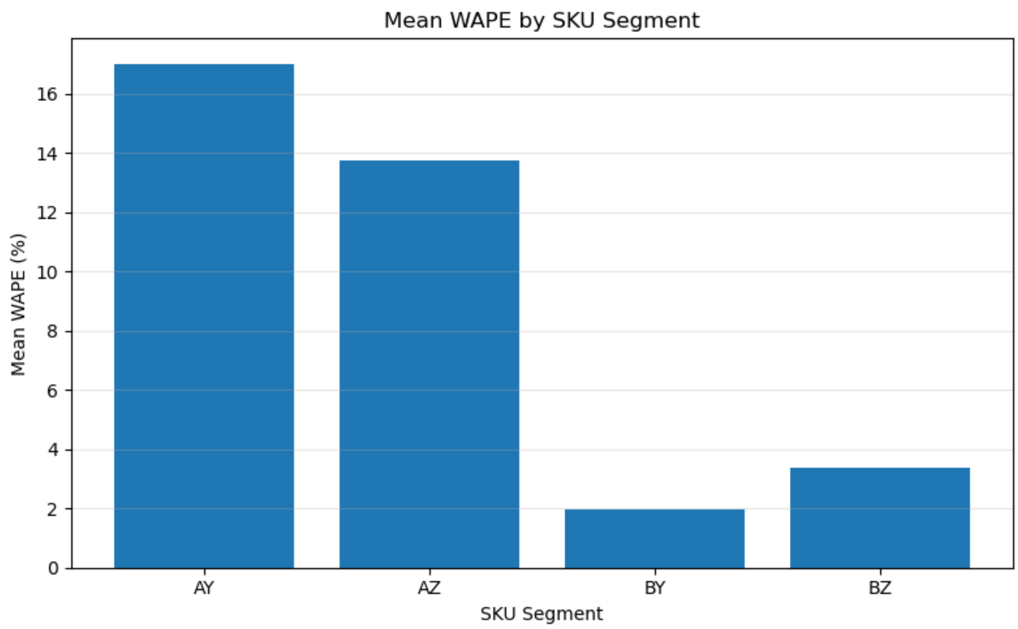

Figure 6.8: Mean WAPE by SKU segment

The following figures illustrate selected SKU-level forecast fits from the 12-week test period. These plots are included to complement the summary metrics and to show how the final forecasts behave at item level.

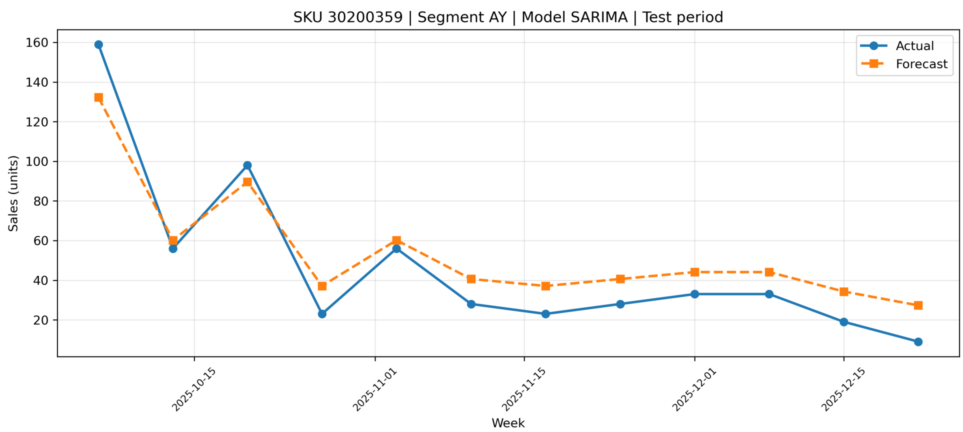

Figure 6.9: Forecast fit for SKU30200359 in segment AY

Figures 6.9 and 6.10 describe examples from the AY segment. In these cases, the forecasts generally follow the direction of the actual series but do not capture every weekly movement exactly. This is expected in short-term SKU-level forecasting, where weekly sales may fluctuate sharply even for products with relatively continuous demand.

Figures 6.11 and 6.12 present examples from the AZ segment. These SKUs are forecasted using SBA, which tends to produce smoother forecasts than the actual weekly observations. This behaviour is consistent with the logic of intermittent-demand forecasting.

Figures 6.13 and 6.14 of BZ segment. Here ,the forecasts are relatively stable, while the actual demand may contain zero-sales weeks or sudden positive observations. This illustrates

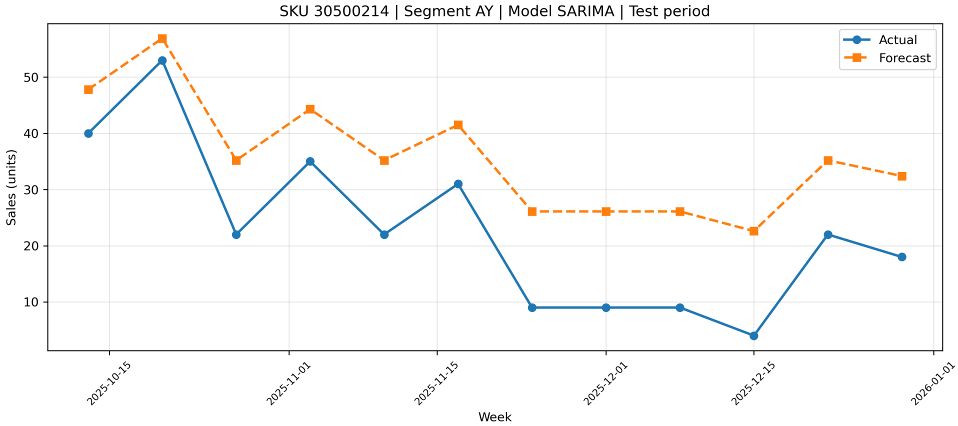

Figure 6.10: Forecast fit for SKU30500214 in segment AY

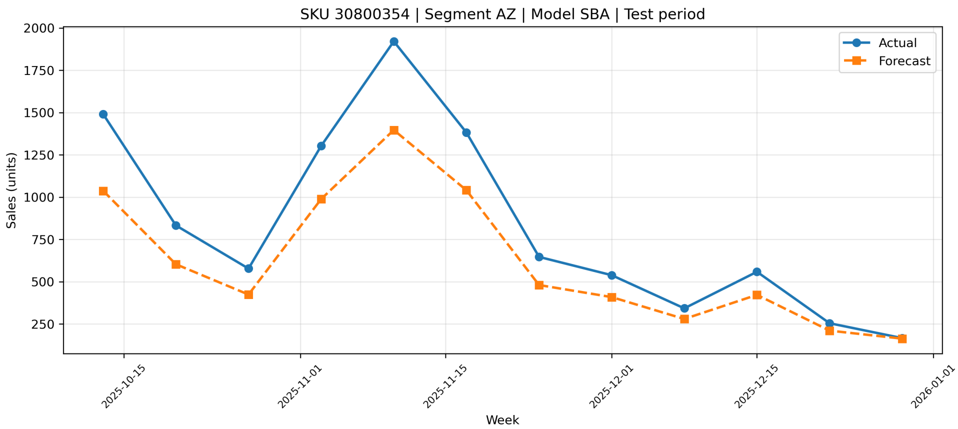

Figure 6.11: Forecast fit for SKU30800354 in segment AZ

the central challenge of intermittent demand: even when average forecast error is low, individual weekly sales events can still be difficult to predict exactly.

Figure 6.15 presents the BY example. Since the final corrected results contain only one BY SKU, this figure should be interpreted as an individual case rather than as evidence for the entire segment.

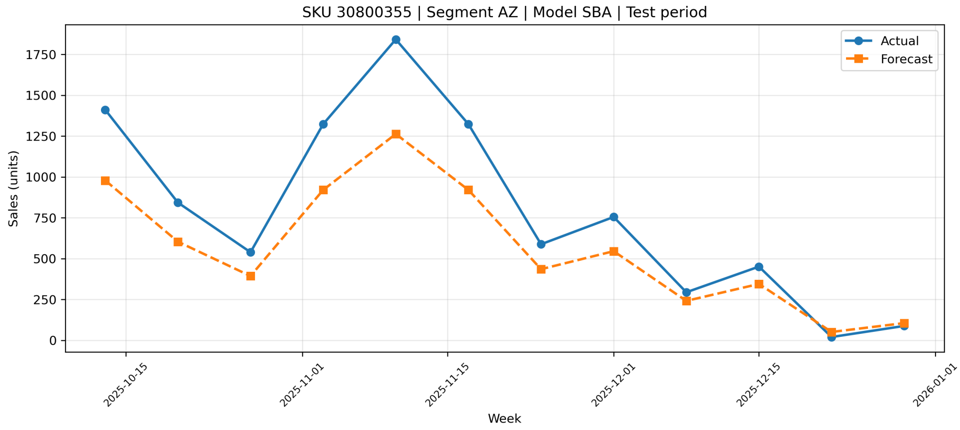

Figure 6.12: Forecast fit for SKU30800355 in segment AZ

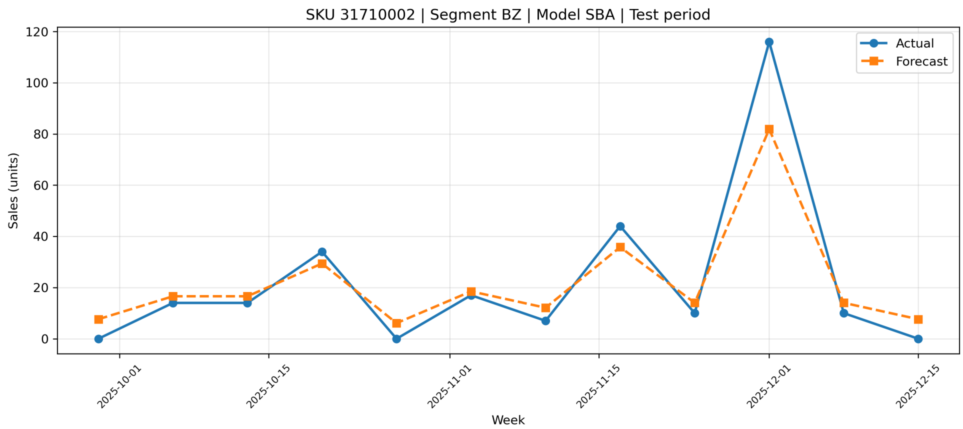

Figure 6.13: Forecast fit for SKU31710002 in segment BZ

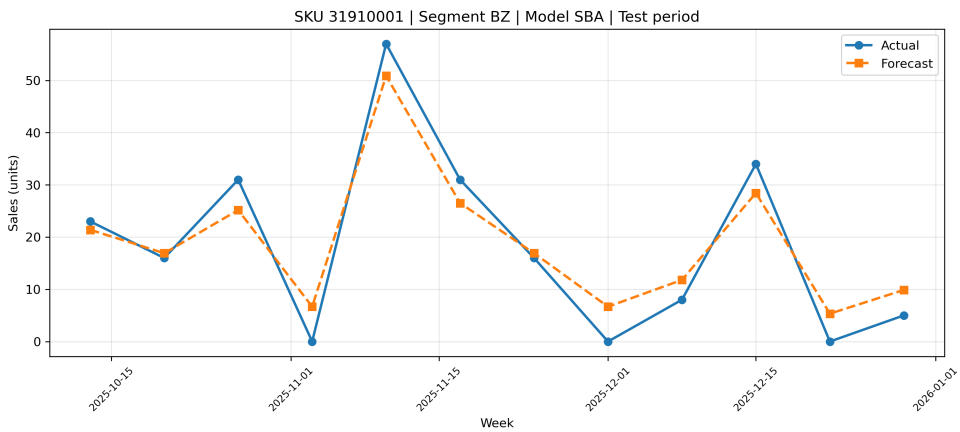

Figure 6.14: Forecast fit for SKU31910001 in segment BZ

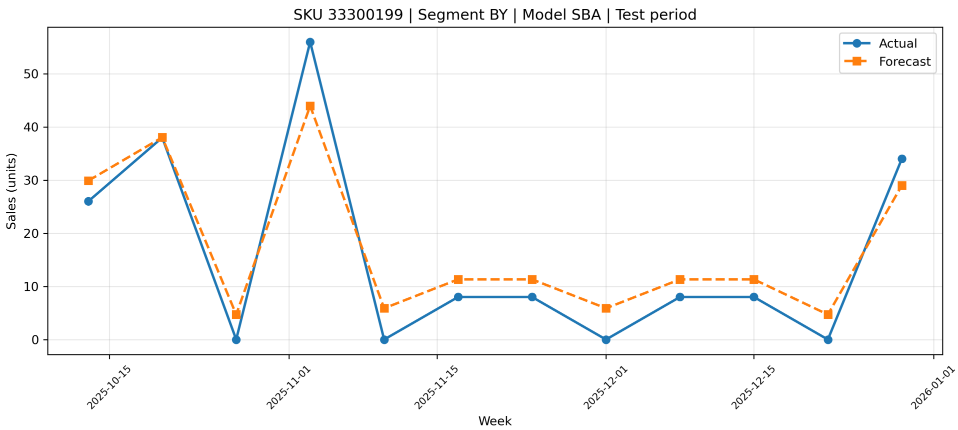

Figure 6.15: Forecast fit for SKU33300199 in segment BY

6.1.7 Interpretation of the BZ Segment Results

The unusually strong performance of the BZ segment is explained by the fact that the selected forecasting method matches the underlying demand structure very well. BZ is an intermittent-demand segment with many zero-sales weeks, and SBA is specifically designed for this type of series. As a result, the model captures the low-frequency demand pattern of the segment more effectively than methods intended for smoother or more regular demand.

In addition, the BZ segment is internally coherent, meaning that the SKUs assigned to it follow a similar sparse-demand profile. This makes the segment-level routing highly effective, because the same method can be applied consistently to products with comparable behaviour. The evaluation is also strictly out of sample, with all predictors built only from past information, so the high accuracy is not due to leakage but to the good match between segment structure and model choice.

6.1.8 Feature Effects and External Variables

The results reported in this chapter are based on the final forecasting pipeline, in which external variables are included only where the selected model allows them. For SARIMA, promotional and holiday variables are incorporated as exogenous regressors. For SBA and Croston, the baseline forecast is adjusted through the external-variable uplift mechanism described in the methodology chapter. This means that the reported model-level and segment-level results already reflect the effect of the external variables.

The results point out that the framework benefits from these adjustments around event weeks, but the main driver of performance remains the underlying demand pattern of each segment. In other words, promotions and holidays help refine the forecast, but they do not fully explain the differences in accuracy across segments. The better results for stable segments such as AY and the strong performance of SBA in several intermittent segments indicate that the choice of forecasting method still matters more than the external variables alone.

6.1.9 Summary of Findings

To sum everything up, there are no single forecasting model is best for all SKU segments. SARIMA performs best for the most stable and seasonal segment, while SBA provides the strongest results for most intermittent or moderately irregular segments. Croston remains useful for very sparse demand patterns, although its performance is less stable than SBA in the final comparison.

The outcomes also rely on the fact that segment structure matters more than any single model choice. MAE and WAPE provide complementary information. Together, they explain that the final segment-based framework produces reasonable forecast accuracy for the evaluated SKUs, while also highlighting the limits of forecasting highly sparse demand.

Chapter 7

Conclusion, Discussion and Recommendations

7.1 Conclusion

7.1.1 Comparison of Baseline Models

The baseline model comparison showed that no single method performs best across all SKU groups. Seasonal Naïve and ETS served as useful benchmark methods, but they were less effective for sparse or intermittent series, especially in zero-heavy segments. SARIMA performed best for the more stable AY segment, where the demand structure is sufficiently regular for seasonal modelling. SBA was selected for several irregular or zero-heavy segments because its bias-corrected intermittent-demand structure matched those patterns better.Lastly, Croston performed best for CX in the baseline model-selection stage.

The final results should not be read as a repetition of the baseline comparison. The baseline table identifies the best-performing candidate model per segment during model selection, whereas the final results report the performance of the selected segment-based forecasting framework on the final evaluated set of SKUs. In the final corrected results, only AY, AZ, BY, and BZ appear. CY and CX remain part of the baseline selection evidence, but they are not included in the final corrected segment-level results.

In order to have a better view on the above, the comparison supports a modular forecasting strategy. The framework does not apply one universal model to all SKUs, but instead routes stable seasonal series to SARIMA and intermittent or sparse series to SBA. This is more appropriate than a one-size-fits-all approach because the portfolio contains both relatively regular and highly irregular demand patterns. The outcomes also show that model performance must be interpreted together with segment composition, since SBA performs strongly because it is well matched to the assigned series.

7.1.2 Interpretation of Results

7.1.2.1 Drivers of Forecast Accuracy Across Segments

The main finding is that forecast accuracy is driven primarily by the underlying demand structure of each segment rather than by model complexity alone. The AY segment that routed to SARIMA, contains SKUs with more continuous and seasonally regular demand histories. These series provide enough temporal structure for SARIMA to capture recurring patterns. On the other hand AY also contains higher-volume products, which naturally produce larger absolute errors even when the relative forecast quality is acceptable. The SBA-routed segments — AZ, BY, and BZ — are characterised by higher zero ratios and more irregular demand arrivals. SBA is designed for exactly this type of demand, since it estimates expected demand by separating demand size from inter-demand intervals. This is why SBA produces lower relative errors in these segments.

The AZ segment deserves special attention. Even though SBA is the selected method for this group, its WAPE remains higher than in BY and BZ. This is consistent with the nature of AZ products: they are operationally important, but their demand is still irregular enough to make forecasting difficult. The result does not indicate that the method is wrong; it shows that this segment is structurally harder to forecast.

7.1.2.2 Interpretation of Forecast Error Metrics

Regarding the two metrics that are used throughout the evaluation (MAE and WAPE). Based on the findings MAE is sensitive to scale, so it tend to be higher for segments with larger sales volumes. Furthermore, WAPE makes it more suitable for comparing segments with different demand levels.

This distinction matters in practice. AY records the highest MAE, but that does not automatically mean it is the worst-performing segment. AY also contains higher-volume SKUs, so larger absolute errors are expected. Its WAPE shows that the forecasts remain operationally usable relative to realised demand. BZ, by contrast, has a much lower MAE, but this partly reflects its lower demand scale. Its WAPE is also low, which is encouraging, but the result should be understood in the context of a segment with frequent zero-demand weeks. In intermittent demand settings, MAE and WAPE should always be interpreted together.

7.1.2.3 Magnitude of Forecast Deviations at the SKU Level

The SKU-level forecast plots in Chapter 6 show the practical size of the deviation between forecasts and actual demand. For AY SKUs using SARIMA, the forecasts generally follow the direction of the actual series but do not match every weekly fluctuation. For SBA-routed segments, forecasts are smoother and deviations arise mainly when actual demand spikes or drops to zero unexpectedly. In BZ, this is especially visible: individual weeks may show actual demand at zero while the forecast remains slightly positive, or the reverse may happen. These deviations are small in absolute terms, but they reflect the limitation of point forecasting for intermittent demand.

7.1.2.4 Systematic Forecast Bias

The two model groups show different bias patterns. For SARIMA-routed SKUs in AY, forecasts tend to be slightly above the actual values in the test period. This can happen because SARIMA estimates its seasonal and trend components from a longer training history and carries that average structure forward into the forecast period. If the test window includes a temporary decline or weaker seasonal activity than the training period, the model will overestimate demand. SARIMA also cannot adapt once the forecast is generated, so it may remain above actual values when the recent demand level drops.

For SBA-routed segments, forecasts tend to be below the actual values. SBA estimates a stable expected demand level from the full training history, including many zero-demand weeks. Because these segments have high zero ratios, the estimated demand level is pulled downward. When the test period contains several positive demand weeks or promotional spikes, the forecast may underestimate realised demand. The uplift adjustment helps in known event weeks, but it cannot fully compensate for unplanned demand bursts or shifts in demand level. In this sense, SBA is conservative by design.

7.1.2.5 Limitations of the BY Segment Results

The BY segment appears in the final results with only one SKU. Although the reported MAE and WAPE are low, they should be interpreted carefully because a single product is not enough to represent an entire demand segment. The result is useful as an individual SKU case, but it should not be generalised to the whole BY class.

This means that the BY outcome is informative only in a limited sense. It shows that the selected forecasting framework can perform well for that particular item, but it does not provide strong evidence about the segment as a whole. Since segment-based forecasting depends on the amount of data available within each group, the BY result should be treated as a case example rather than a robust segment-level conclusion.

7.1.3 Limitations of the Study

Several limitations should be acknowledged. First, the forecasting pipeline is evaluated on a single 12-week test horizon per SKU, which limits the generalisability of the accuracy estimates. A rolling-window or cross-validation approach would provide a more robust assessment, but it would also require more computation and a longer evaluation setup. Second, the promotional uplift factors are estimated from training data and applied multiplicatively to the base forecast.

This assumes that future promotional effects are similar to historical ones, which may not always hold if the promotion strategy changes.

Third, the exclusion of the CZ segment means that the framework does not cover the full SKU portfolio. Extremely sparse or low-volume products remain outside the final forecasting pipeline. Finally, the BY segment is represented by only one SKU in the final results, which limits the strength of any segment-level conclusion for that class.

7.2 Discussion

The results show that a segment-aware forecasting framework is a viable approach for short-term demand planning in a retail setting. The key insight is that demand heterogeneity across the product portfolio should guide model selection. By aligning each method with the demand type it fits best — SARIMA for stable seasonal series and SBA for intermittent or sparse series — the framework achieves reasonable accuracy across a diverse set of SKU groups without relying on a single universal model.

The role of promotional and holiday features is also relevant. These variables improve the final forecasts, but their role is secondary to the baseline demand structure. For SARIMA, they enter as exogenous regressors and influence the model directly. For SBA and Croston, they are used as post-hoc adjustments based on historical event patterns. In both cases, the event-based effects help around known promotional or holiday weeks, but they cannot compensate for weak or unstable historical demand.

The fallback mechanism is also important. When minimum history or non-zero demand requirements are not met, the pipeline uses a last-value naïve forecast. This keeps the system operational and ensures non-negative outputs, while also making clear which forecasts are based on a fallback rather than on the selected model. In practical terms, that matters because newly listed or very sparse products often do not provide enough history for reliable model fitting.

7.3 Recommendations

Based on the results and limitations of the study, the following recommendations are proposed for future work and operational use:

1. **Extend the evaluation to a rolling-window framework.** A single 12-week test split is sufficient for thesis-level evaluation, but operational deployment would benefit from a rolling or cross-validated evaluation across multiple historical windows. This would provide more stable accuracy estimates and reduce sensitivity to the specific test period.

2. **Expand the treatment of the CZ segment.** Extremely sparse products are currently excluded from the forecasting pipeline. Future work could explore aggregation-based ap-

proaches, such as forecasting at category level, or intermittent-demand methods that are more robust to near-zero histories.

3. **Refine the promotional uplift estimation.** The current uplift approach relies on simple ratio estimation from historical training data. A more detailed approach could model the promotional effect by promotion type, duration, and intensity, so that the uplift is estimated more precisely.

4. **Incorporate inventory signals.** The current framework forecasts demand independently of stock availability. Including stockout information would improve the quality of the training data and reduce the risk of learning distorted demand patterns.

5. **Update segmentation periodically.** The ABC–XYZ segmentation is computed once from the historical data. In practice, demand patterns can change over time, so periodic re-segmentation would help keep the routing logic aligned with current behaviour.

6. **Evaluate machine learning alternatives for stable segments.** Although SARIMA performs well for AY, this segment may also be suitable for models that combine multiple exogenous inputs with the sales series. Such alternatives could be assessed in future work as a comparison point, but the current study already shows that the chosen classical methods are effective for the observed demand structure.

Bibliography

Abolghasemi, Mahdi, Jason Hurley, Ali Eshragh, and Behnam Fahimnia (2020). "Demand Forecasting in the Presence of Systematic Events: Cases in Capturing Sales Promotions." In: *International Journal of Production Economics* 230, p. 107892. DOI: 10.1016/j.ijpe.2020.107892.

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo (2018). "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction." In: *Computational Statistics & Data Analysis* 120, pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

Bojer, Casper Solheim, Iskra Dukovska-Popovska, Flemming Christensen, and Kenn Steger-Jensen (2019). "Retail Promotion Forecasting: A Comparison of Modern Approaches." In: *Advances in Production Management Systems. Towards Smart Production Management Systems*. IFIP Advances in Information and Communication Technology. Springer, pp. 575–582. DOI: 10.1007/978-3-030-29996-5_66.

Croston, J. D. (1972). "Forecasting and Stock Control for Intermittent Demands." In: *Operational Research Quarterly* 23.3, pp. 289–303. DOI: 10.2307/3007885.

Fildes, Robert, Shaohui Ma, and Stephan Kolassa (2022). "Retail Forecasting: Research and Practice." In: *International Journal of Forecasting* 38.4, pp. 1283–1318. DOI: 10.1016/j.ijforecast.2019.06.004.

Flores, Benito E. and D. Clay Whybark (1986). "Multiple Criteria ABC Analysis." In: *International Journal of Operations & Production Management* 6.3, pp. 38–46. DOI: 10.1108/eb054765.

Huber, Jakob and Heiner Stuckenschmidt (2020). "Daily Retail Demand Forecasting Using Machine Learning with Emphasis on Calendric Special Days." In: *International Journal of Forecasting* 36.4, pp. 1420–1438. DOI: 10.1016/j.ijforecast.2020.02.005.

Hyndman, Rob J. and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. 3rd ed. Melbourne, Australia: OTexts. URL: https://otexts.com/fpp3/.

Hyndman, Rob J. and Yeasmin Khandakar (2008). "Automatic Time Series Forecasting: The forecast Package for R." In: *Journal of Statistical Software* 27.3, pp. 1–22. DOI: 10.18637/jss.v027.i03.

Hyndman, Rob J. and Anne B. Koehler (2006). "Another Look at Measures of Forecast Accuracy." In: *International Journal of Forecasting* 22.4, pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Hyndman, Rob J., Anne B. Koehler, Ralph D. Snyder, and Simone Grose (2002). "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods." In: *International Journal of Forecasting* 18.3, pp. 439–454. DOI: 10.1016/S0169-2070(01)00110-8.

Kaya, Gamze Ogcu, Merve Şahin, and Ömer Fahrettin Demirel (2020). "Intermittent Demand Forecasting: A Guideline for Method Selection." In: *Sadhana* 45.51, pp. 1–7. DOI: 10.1007/s12046-020-1285-8.

Kolassa, Stephan and Wolfgang Schütz (2007). "Advantages of the MAD/Mean Ratio over the MAPE." In: *Foresight: The International Journal of Applied Forecasting* 6, pp. 40–43.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2022). "M5 Accuracy Competition: Results, Findings, and Conclusions." In: *International Journal of Forecasting* 38.4, pp. 1346–1364. DOI: 10.1016/j.ijforecast.2021.11.013.

Syntetos, Aris A. and John E. Boylan (2005). "The Accuracy of Intermittent Demand Estimates." In: *International Journal of Forecasting* 21.2, pp. 303–314. DOI: 10.1016/j.ijforecast.2004.10.001.

Syntetos, Aris A., John E. Boylan, and J. D. Croston (2005). "On the Categorization of Demand Patterns." In: *Journal of the Operational Research Society* 56.5, pp. 495–503. DOI: 10.1057/palgrave.jors.2601841.

Willemain, Thomas R., Charles N. Smart, and Henry F. Schwarz (1994). "Forecasting Intermittent Demand in Manufacturing: A Comparative Evaluation of Croston's Method." In: *International Journal of Forecasting* 10.4, pp. 529–538. DOI: 10.1016/0169-2070(94)90021-3.