A Comparative Study of Classical Time Series and Machine Learning Models for Retail Demand Forecasting

Student Number: 249170797

Student Name: Manikanta Konkathi

Supervisor: Lukasz Piwek

MSc in Business Analytics Dissertation

Date: 01/09/2025

Submitted as a part of the requirement for completing an MSc in Business Analytics at the University of Bath

Word Count: 14952

SCHOOL OF MANAGEMENT

UNIVERSITY OF BATH

SEPTEMBER 2025

Acknowledgements

I would like to express my sincere gratitude to **Dr Lukasz Piwek** for his exceptional supervision throughout this project. His clear guidance, thoughtful critique, and unfailing encouragement were invaluable at every stage—from shaping the research question to refining the analysis and presentation of results. I am especially grateful for his patient feedback on successive drafts and for the high standards he set for methodological rigor. Any remaining errors are my own. I am equally indebted to my family for their unwavering support, patience, and belief in me; their understanding made the long hours of study and writing possible.

Abstract

This research paper presents a rigorous comparative analysis of classical time series models and machine learning approaches for forecasting weekly retail sales at a chain level using real-world data. Focusing on six models—Random Forest, XGBoost, Prophet, Seasonal Naïve, SARIMA, and ETS—we implement a robust Monte Carlo cross-validation scheme with consistent feature sets across all methods. Our findings demonstrate that machine learning models, particularly Random Forest and XGBoost, significantly outperform classical benchmarks with substantial reductions in forecasting error metrics, including a 38.3% improvement in MAPE over the best classical method. Prophet, treated as a hybrid machine learning approach, provides a valuable balance between interpretability and accuracy. Feature importance analyses reveal that lagged sales and seasonal indicators are critical drivers of forecasting performance, underscoring the importance of sophisticated feature engineering. The study addresses key gaps in the literature regarding fair model comparison frameworks, evaluation rigor, and practical deployment considerations. Results translate into meaningful economic value for retail operations, justifying the adoption of machine learning forecasting systems. This work contributes both theoretical insights and actionable guidance for advancing retail demand forecasting, advocating a paradigm shift from classical methods toward data-driven, machine learning-enabled frameworks.

Table of Contents

List of Figures

List of Tables

1.INTRODUCTION

1.1 Industry Context and Motivation

The global retail industry, valued at approximately £1.1 trillion annually, with the UK market accounting for around £490 billion in 2022 (Statista, 2023), depends critically on accurate demand forecasting to optimise inventory management, supply chain efficiency, and customer satisfaction. Forecasting errors impose substantial costs through excess inventory carrying charges and lost sales resulting from stockouts. Industry analyses suggest that improvements in forecasting accuracy can reduce operational costs by 10–15% while simultaneously enhancing service levels (McKinsey, 2024). Despite decades of advancements in statistical forecasting methodologies, retail organisations continue to face notable challenges in achieving accurate predictions, particularly during highly variable promotional periods and seasonal demand fluctuations characteristic of modern retail environments.

1.2 Forecasting Methods and Research Gaps

Accurate demand forecasting is a critical function in retail, directly influencing inventory management, supply chain efficiency, and profitability (Nasseri, M. et al., 2023). Retailers like Walmart deal with complex sales patterns affected by seasonality (e.g. holidays), trends, and external factors (promotions, economic indicators). Classical statistical models—such as ARIMA and Exponential Smoothing—have traditionally formed the core forecasting toolset, valued for their theoretical rigor and interpretability. However, these classical methods often struggle to represent nonlinear sales dynamics and to incorporate multiple external regressors simultaneously (Kontopoulou et al., 2023). Recently, machine learning (ML) methods, particularly tree-based ensemble algorithms like Random Forest and gradient boosting (XGBoost), alongside hybrid statistical-ML frameworks like Prophet, have gained prominence. These approaches offer enhanced flexibility to model complex relationships across diverse features and have exhibited superior predictive performance in various domains including retail (Nasseri, M. et al., 2023; Makridakis et al., 2022).

However, I envisaged key challenges in establishing a fair comparison: ensuring identical feature access, adopting rigorous cross-validation, and reconciling mixed empirical findings in the literature. Academic studies reveal persistent ambiguities regarding the circumstances under which ML methods decisively outperform classical approaches. While benchmark competitions such as M4 and M5 highlight considerable accuracy gains from ML and hybrid methods, many empirical studies report mixed results. Moreover, fragmented evaluation protocols—characterised by the use of proprietary datasets, limited benchmarking against strong classical baselines, and insufficient validation rigour—complicate efforts to generalise findings or ascertain ML's true practical advantage (Kontopoulou et al., 2023). The integration of external factors like promotions and prices, well known to enhance forecasting accuracy, is unevenly addressed, often favouring ML methods by default due to their inherent flexibility. Furthermore, cross-validation techniques specifically tailored to time series forecasting contexts are not consistently applied, creating risks of overfitting and biased performance estimates.

1.3 Research Question and Objectives

Research Question:
Given these gaps, this study seeks to answer: **How does the forecasting performance of classical time series models compare to machine learning models in predicting retail sales demand?** In particular, we examine whether machine learning models enhanced with lagged sales and seasonal features can outperform classical benchmarks when forecasting weekly sales aggregated at the retail chain level.

To address this question, the study objectives are threefold:

- **Implement multiple forecasting models** from both classical time series methods (Seasonal Naïve, ARIMA/SARIMA, exponential smoothing) and machine learning methods (Random Forest, XGBoost gradient boosting, Prophet) using the same real-world retail dataset.
- **Compare their predictive performance** on out-of-sample forecasts using consistent accuracy metrics. We use a publicly available Walmart weekly sales dataset, which includes exogenous features (holiday flags, fuel prices, CPI, etc.), ensuring a realistic evaluation.
- **Evaluate forecast accuracy** with common error metrics – specifically Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) – under a robust cross-validation scheme. We employ a **Monte Carlo cross-validation** with multiple train/test splits to assess how model performance varies over different periods.

1.4 Research Contributions and Significance:

This investigation advances academic knowledge and practical application across multiple dimensions. Theoretically, it challenges traditional assumptions about the sufficiency of linear time series models in complex retail settings, offering empirical evidence for a paradigm shift towards machine learning approaches that better capture nonlinearities and multivariate dependencies. Methodologically, the research contributes a replicable evaluation framework that rigorously addresses temporal dependence and feature parity, addressing critical evaluation limitations highlighted in prior literature. Practically, it quantifies performance differences in terms directly translatable to business impact, providing retail organisations with evidence-based guidance for forecasting system investments and contextualizing where classical methods may remain appropriate.

2. LITERATURE REVIEW: THE EVOLUTION FROM CLASSICAL TO MACHINE LEARNING APPROACHES IN RETAIL DEMAND FORECASTING

The quest for accurate demand forecasting has fundamentally transformed over the past two decades, driven by a paradigmatic shift from classical statistical methods toward machine learning approaches that promise to capture the complexity of modern retail environments. While traditional time series models like ARIMA and exponential smoothing dominated forecasting practice for decades due to their mathematical elegance and interpretability, the emergence of volatile consumer behavior, high-frequency data availability, and computational advances has exposed critical limitations in these classical frameworks. This evolution represents more than a simple methodological upgrade—it embodies a fundamental reconceptualization of forecasting from linear, univariate modeling toward nonlinear, multivariate learning systems capable of integrating diverse data sources and adapting to complex retail dynamics.

The central tension driving contemporary forecasting research lies in reconciling the proven reliability of classical methods with the transformative potential of machine learning approaches. Recent benchmarking studies, particularly the M5 Forecasting Competition focused on Walmart retail data, have demonstrated that machine learning methods can achieve substantial accuracy improvements over traditional statistical approaches—yet this superiority remains context-dependent and theoretically underexplored (Makridakis et al., 2022). This literature review synthesizes the evolutionary trajectory of retail demand forecasting research, examining how foundational statistical theories have been challenged, extended, and sometimes validated by machine learning innovations, while identifying critical gaps that continue to limit both theoretical understanding and practical implementation.

2.1 The foundation and enduring relevance of classical forecasting

Classical time series forecasting emerged from decades of statistical theory development, with ARIMA models and exponential smoothing methods establishing the mathematical foundation for decomposing retail demand into trend, seasonal, and irregular components. These methods gained widespread adoption not merely due to their mathematical rigor, but because they addressed fundamental business needs: *interpretability for decision-makers, robustness with limited data, and computational efficiency for large-scale implementations* (Hyndman & Athanasopoulos, 2018). The theoretical elegance of these approaches lies in their explicit modeling of time series components through parsimonious parameter structures that business practitioners can understand and validate against domain knowledge.

The empirical performance of classical methods established early benchmarks that continue to influence contemporary research. ARIMA models, with their ability to capture autocorrelation patterns through autoregressive and moving average terms, proved particularly effective for retail categories with stable seasonal patterns and linear trends. Brykin (2024) demonstrated this enduring strength, finding that basic ARIMA models provided consistent accuracy across multiple retail categories, effectively capturing overall trends and seasonality with minimal parameter tuning. Similarly, exponential smoothing methods, particularly the Holt-Winters triple smoothing approach, achieved remarkable success in industry applications due to their automatic model selection capabilities and recursive updating equations that naturally weight recent observations more heavily. Many retail studies have used ARIMA as a baseline; for instance, *Xian et al. (2023)* compared SARIMA with Holt-Winters and ETS (Exponential Smoothing state-space model) for disease incidence forecasting, finding that a well-tuned Holt-Winters exponential smoothing model slightly outperformed SARIMA and ETS on their monthly series. This suggests that for data with strong seasonal patterns and a relatively stable trend, exponential smoothing methods can be very competitive. ETS and SARIMA both proved effective for short-term retail sales forecasts in prior studies (İnce & Taşdemir, 2024), though these classical models may struggle when patterns become complex or when there are abrupt changes (e.g., due to promotions).

The persistent relevance of classical methods extends beyond their historical importance to their continued competitive performance in specific contexts. Early forecasting competitions (M1-M3) demonstrated that ARIMA and exponential smoothing were remarkably difficult to outperform, particularly for univariate time series with clear patterns and sufficient historical data. *This resilience stems from their fundamental strength: classical methods excel when the underlying assumptions of linear relationships and stationary patterns hold, requiring minimal data preprocessing and providing transparent parameter interpretation* that facilitates business understanding and trust. However, classical models have limitations. They generally assume a specific data generating process (linear combinations of past values and errors for ARIMA; exponentially decaying weights for ETS). Thus, they *struggle with nonlinear effects* and interactions that often occur in retail (e.g. a promotion might spike sales in a nonlinear way). They also typically handle exogenous regressors in a linear fashion. While SARIMA and ETS can include external regressors (SARIMAX, ETSX), this requires careful manual specification, and the effects are assumed additive and fixed over time. In practice, not all academic works include these exogenous factors. As a result, baseline models in literature may be *under-fitting* the true drivers of demand. For example, *Kontopoulou et al. (2023)* note in their review that many studies benchmark ARIMA on datasets without integrating additional predictors, whereas machine learning models in those studies often use extra features – leading to an unfair comparison (Kontopoulou et al., 2023). Our study addresses this by giving each model access to the same feature set (where possible) to enable a fair head-to-head evaluation.

The Prophet framework, developed by Taylor and Letham (2018), represents a crucial bridge between classical statistical thinking and modern computational approaches. By combining piecewise linear trends with Fourier-based seasonal terms and explicit holiday effects within a Bayesian framework, Prophet demonstrates how classical decomposition principles can be enhanced through contemporary statistical computing. *Žunić et al. (2020) demonstrated that Facebook's Prophet—by incorporating user-defined holiday and promotion indicators alongside its piecewise linear trend and Fourier seasonality components—significantly improved retail sales forecasts on real-world store data, highlighting the value of embedding domain knowledge in the model***.** However, comparative studies reveal that Prophet's performance varies significantly across product categories, excelling in scenarios with clear seasonal patterns while struggling with highly volatile or irregular demand series (Brykin, 2024).

2.2 The machine learning revolution in forecasting

The emergence of machine learning approaches in forecasting represents a fundamental philosophical shift from parametric modeling of time series components toward algorithmic pattern discovery in high-dimensional feature spaces. This transformation was initially driven by the recognition that retail demand patterns often exhibit nonlinear relationships, cross-product interactions, and complex dependencies on external factors that classical linear models struggle to capture. Tree-based ensemble methods, particularly Random Forest and gradient boosting algorithms like XGBoost and LightGBM, pioneered this evolution by demonstrating superior performance when rich feature sets including price, promotions, calendar effects, and store metadata became available.

In retail forecasting, *tree-based ensemble models* have gained traction due to their ability to naturally incorporate many input features (lags, exogenous variables, categorical indicators) and handle large datasets. Hasan et al. (2022) performed a comparative analysis on the M5 Walmart dataset, demonstrating that Random Forest—by leveraging holiday and store-specific indicator features—consistently achieved lower RMSE and MAPE than ARIMA models, owing to its ability to capture nonlinear interactions that static linear models miss**.** *Gradient Boosting Machines* like *XGBoost* have also been applied successfully. For example, Nasseri, M. et al. (2023) compare tree-based ensembles to an LSTM on grocery sales and find that ensemble methods (including Extra Trees, RF, XGBoost, and Gradient Boosting) achieved similar performance to the LSTM, with the best ensemble slightly outperforming the neural network in terms of MAE and MAPE. The M5 Forecasting Competition marked a decisive turning point in the classical versus machine learning debate. All top 50 methods in this Walmart-focused competition employed machine learning approaches, predominantly LightGBM variants, achieving over 20% accuracy improvements compared to the best statistical benchmarks (Makridakis et al., 2022). This dramatic performance gap emerged from machine learning's ability to exploit the hierarchical structure of retail data, learning correlations across thousands of products and stores simultaneously while automatically discovering feature interactions that would be impossible to specify manually in classical models. LightGBM's particular success stemmed from its computational efficiency and ability to handle the large-scale, sparse data characteristics typical of retail forecasting problems.

Deep learning methods, especially Long Short-Term Memory (LSTM) networks, introduced another dimension to the machine learning revolution by addressing the temporal dependencies that tree-based methods handle less naturally. LSTMs excel at modeling complex nonlinear dynamics and long-range dependencies in time series, making them particularly suitable for capturing subtle patterns such as promotional lift effects and seasonal variations that evolve over time. Ahmadov and Helo (2023) demonstrated this capability in e-commerce contexts, where deep neural networks achieved up to 35% lower error rates than classical methods on intermittent demand data by better capturing the irregular, "nervous" patterns characteristic of online sales.

However, the machine learning revolution has not followed a simple trajectory toward universal superiority. **The performance of deep learning methods proves highly sensitive to hyperparameter tuning, data quality, and architectural choices, leading to significant variability across implementations and datasets** (Brykin, 2024). This sensitivity, combined with the substantial computational resources and expertise required for effective deployment, has led many researchers to explore hybrid approaches that combine the interpretability and stability of classical methods with the pattern recognition capabilities of machine learning algorithms. **Jahin et al. (2024)** propose an explainable multi-channel deep neural network that fuses various data sources (sales history, weather, Google Trends, etc.) for supply chain demand forecasting. Their approach, while complex, achieved high accuracy and provided interpretability through attention mechanisms, indicating the potential of modern deep learning if coupled with feature engineering and interpretability techniques.

The evolution toward ensemble and hybrid methods represents a sophisticated response to the recognition that no single approach dominates across all forecasting contexts. The winning method of the M4 competition exemplified this strategy by combining exponential smoothing with recurrent neural networks, achieving superior accuracy through complementary strengths rather than replacing one method with another (Smyl, 2020). This hybrid philosophy has become increasingly prevalent in retail applications, where practitioners commonly ensemble ARIMA, tree-based methods, and neural networks to improve robustness while hedging against individual model weaknesses.

2.3 Retail context and real-world validation

The application of forecasting methods in retail contexts reveals critical insights about the interaction between algorithmic sophistication and business reality that pure methodological studies often overlook. Retail demand forecasting operates within complex ecosystems where promotional activities, seasonal events, competitor actions, and macroeconomic factors create multifaceted pattern structures that challenge both classical and machine learning approaches in different ways. The Walmart M5 competition dataset has become a crucial testbed precisely because it captures this complexity through real-world retail data spanning multiple years, thousands of SKUs, and diverse geographical locations.

Recent industry applications demonstrate that the choice between classical and machine learning methods depends critically on organizational capabilities and business constraints beyond pure forecasting accuracy. İnce and Taşdemir (2024) illustrated this complexity in their analysis of U.S. furniture store sales, where a multiple linear regression model incorporating external factors achieved slightly better performance (3.47% MAPE) than Holt-Winters smoothing (4.21% MAPE), but required substantially more data preparation and domain expertise. This finding highlights a recurring theme in retail forecasting: incremental accuracy improvements from sophisticated methods must be weighed against implementation complexity, interpretability requirements, and organizational change management costs.

Intermittent demand patterns, characteristic of many retail categories, present particular challenges that reveal the limitations of both classical and machine learning approaches. Traditional methods like Croston's approach were specifically designed for sparse demand patterns but often fail to capture the complex drivers of demand occurrence versus magnitude. Machine learning methods can potentially learn these patterns from data but require careful handling of zero-inflation and may overfit to noise in sparse series. The M5 competition results suggested that globally trained machine learning models could outperform specialized intermittent demand methods by borrowing strength across related products, but this finding requires validation across diverse retail contexts beyond the Walmart ecosystem.

2.4 Comparative studies and methodological insights

The proliferation of comparative studies between classical and machine learning forecasting methods has yielded nuanced insights that challenge simplistic narratives about machine learning superiority while highlighting the context-dependent nature of forecasting performance. Systematic comparisons reveal that method superiority depends critically on data characteristics, forecast horizon, evaluation metrics, and business constraints rather than following universal patterns (Kontopoulou et al., 2023). This complexity has led researchers to develop more sophisticated evaluation frameworks that move beyond single-metric comparisons toward multidimensional performance assessment.

Recent comparative research emphasizes the importance of evaluation methodology in drawing valid conclusions about relative method performance. The adoption of time-series cross-validation techniques, which simulate operational forecasting through rolling window evaluation, has revealed that many early claims about machine learning superiority were based on inadequate validation procedures that failed to account for temporal dependencies and overfitting risks (Makridakis et al., 2022). *Rigorous evaluation protocols now require multiple validation windows, statistical significance testing, and consideration of forecast horizon effects to establish reliable performance comparisons.*

A recent systematic review by *Hall & Rasheed (2025)* surveyed **79** time-series papers (including several retail tasks) and found that *tree-based boosting models (XGBoost/LightGBM/CatBoost)* slightly *outperformed deep learning on average* and were *dramatically faster to train—with a mean training-time advantage of 126,934.94%* across studies—while maintaining strong accuracy. These results align with **M5** competition evidence where tree-boosting methods outperformed statistical baselines on Walmart's hierarchical retail series, underscoring that evaluation conclusions depend strongly on data richness and the validation design. A comparable finding is reported by Hobor et al. (2025), who conducted an exhaustive evaluation on a high-resolution brick-and-mortar retail dataset. They compared tree-based ensembles (including LightGBM) against state-of-the-art neural architectures (such as N BEATS and NHITS, which capture long-term dependencies similarly to LSTM). Across multiple experimental settings, LightGBM consistently delivered superior forecasting accuracy and required substantially less training time than the recurrent neural models, demonstrating that well-tuned tree-based methods can outperform deep learning approaches in complex retail forecasting tasks. Comparative findings emphasize not only accuracy but also scalability and computational efficiency (Petropoulos et al., 2021).

These findings point toward a fundamental insight: the evolution from classical to machine learning approaches represents not a replacement but an expansion of the forecasting toolkit, *where method selection should be guided by data characteristics and business context rather than algorithmic sophistication.* Brykin's (2024) systematic comparison across multiple retail categories exemplifies this perspective, showing that ARIMA models remained highly competitive for daily and monthly forecasts with clear patterns, while LSTM networks excelled primarily in high-frequency scenarios with complex nonlinear relationships, and Prophet provided reliable performance when explicit incorporation of calendar effects was crucial.

2.5 Research gaps and future directions

Despite the extensive use of classical time series methods in retail demand forecasting, recent research has identified limitations in their ability to model complex, nonlinear, and high-frequency sales patterns, especially in large-scale retail datasets (Makridakis et al., 2018). Concurrently, while machine learning (ML) models have shown promising results in various domains, their application to retail sales forecasting—particularly in systematic comparisons against traditional methods using real-world datasets—remains relatively underexplored and fragmented in the literature (Oreshkin et al. 2020;Rabenoro et al., 2022).

Despite substantial progress in retail forecasting research, critical gaps remain that limit both theoretical understanding and practical implementation of comparative approaches between classical and machine learning methods. *The most significant gap lies in the absence of systematic frameworks for method selection based on data characteristics and business context.* While numerous studies demonstrate that no single approach dominates universally, little guidance exists for practitioners attempting to choose appropriate methods for specific retail forecasting challenges. This gap is particularly problematic given the resource requirements and expertise needed for implementing sophisticated machine learning approaches compared to classical alternatives.

The evaluation and benchmarking infrastructure for retail forecasting research remains fragmented and inconsistent. Kontopoulou et al. (2023) highlight that many studies propose new methods without proper comparison against both classical baselines and state-of-the-art machine learning approaches, making it difficult to assess genuine contributions to the field. The lack of standardized evaluation protocols, consistent datasets, and agreed-upon performance metrics creates a fragmented research landscape where conflicting claims about method superiority persist without resolution. The success of the M5 competition demonstrates the value of common benchmarking platforms, but broader adoption of such standardized evaluation frameworks remains limited.

The integration of external factors and contextual variables represents another underexplored area with significant practical implications. While the M5 competition included price and calendar data that top teams leveraged effectively, most academic studies continue to focus on univariate time series modeling due to data availability constraints and analytical simplicity. Real-world retail forecasting accuracy could be substantially improved through systematic exploration of promotional effects, weather patterns, economic indicators, and competitive dynamics, but research on optimal feature engineering and selection strategies remains sparse. This gap is particularly important because machine learning methods' superior performance often depends on access to rich feature sets that classical methods cannot naturally accommodate.

The interpretability and explainability of machine learning forecasting models presents both theoretical and practical challenges that current research has barely begun to address. As machine learning methods gain adoption in retail planning processes, the need for transparent, explainable predictions becomes critical for decision-maker acceptance and regulatory compliance. Existing interpretability techniques developed for tabular machine learning problems may not translate effectively to time series forecasting contexts, where temporal dependencies and hierarchical structures create additional complexity layers. Research on developing forecasting-specific interpretability methods that can explain both individual predictions and systematic patterns in model behavior represents a crucial frontier for practical adoption.

The research trajectory reveals several key insights that should guide future investigations. First, the performance comparison between classical and machine learning methods is fundamentally context-dependent, with method superiority varying across product categories, forecast horizons, and data characteristics in ways that current theory cannot fully predict. Second, hybrid and ensemble approaches that combine classical and machine learning methods often outperform individual approaches, suggesting that the future lies in intelligent method combination rather than wholesale replacement. Third, the evaluation infrastructure for comparative forecasting research requires substantial improvement to enable reliable conclusions about method performance and theoretical understanding.

The path forward demands research that bridges the gap between algorithmic innovation and practical implementation, focusing on systematic method selection frameworks, standardized evaluation protocols, and interpretability solutions that enable widespread adoption of advanced forecasting approaches in retail contexts. The ultimate goal should not be identifying a single superior method but developing principled approaches for matching forecasting techniques to specific retail challenges while maintaining the transparency and reliability that business decision-making requires. This balanced perspective acknowledges both the transformative potential of machine learning approaches and the enduring value of classical methods, positioning future research to advance practical forecasting capabilities rather than pursuing algorithmic sophistication for its own sake.

3.METHODOLOGY

3.1 Chapter Overview

This methodology chapter outlines the comprehensive approach employed to compare classical time series forecasting models with machine learning approaches for retail demand forecasting using Walmart's historical sales data. The research adopts a rigorous time-aware validation framework designed to evaluate model performance whilst preserving the temporal dependencies inherent in weekly retail sales patterns. The analytical pipeline encompasses five distinct stages: data preprocessing and aggregation to chain level, exploratory data analysis to inform feature engineering decisions, systematic feature creation incorporating temporal lags and seasonal components, Monte Carlo time-series cross-validation for robust model assessment, and comparative evaluation across multiple forecasting paradigms including SARIMA, ETS, seasonal naïve, Prophet, Random Forest and XGBoost methods.

The methodology is grounded in the principle that retail demand forecasting requires models capable of capturing both short-term autocorrelations and long-term seasonal patterns whilst accommodating the non-stationary characteristics typical of business time series. By aggregating individual store-department combinations to chain level, this initial analysis phase establishes baseline performance benchmarks before progressing to more granular forecasting challenges. The evaluation framework employs contiguous block cross-validation to ensure temporal integrity, with model performance assessed using standard forecasting accuracy metrics (MAE, RMSE, MAPE) computed across multiple holdout periods to provide robust statistical inference.

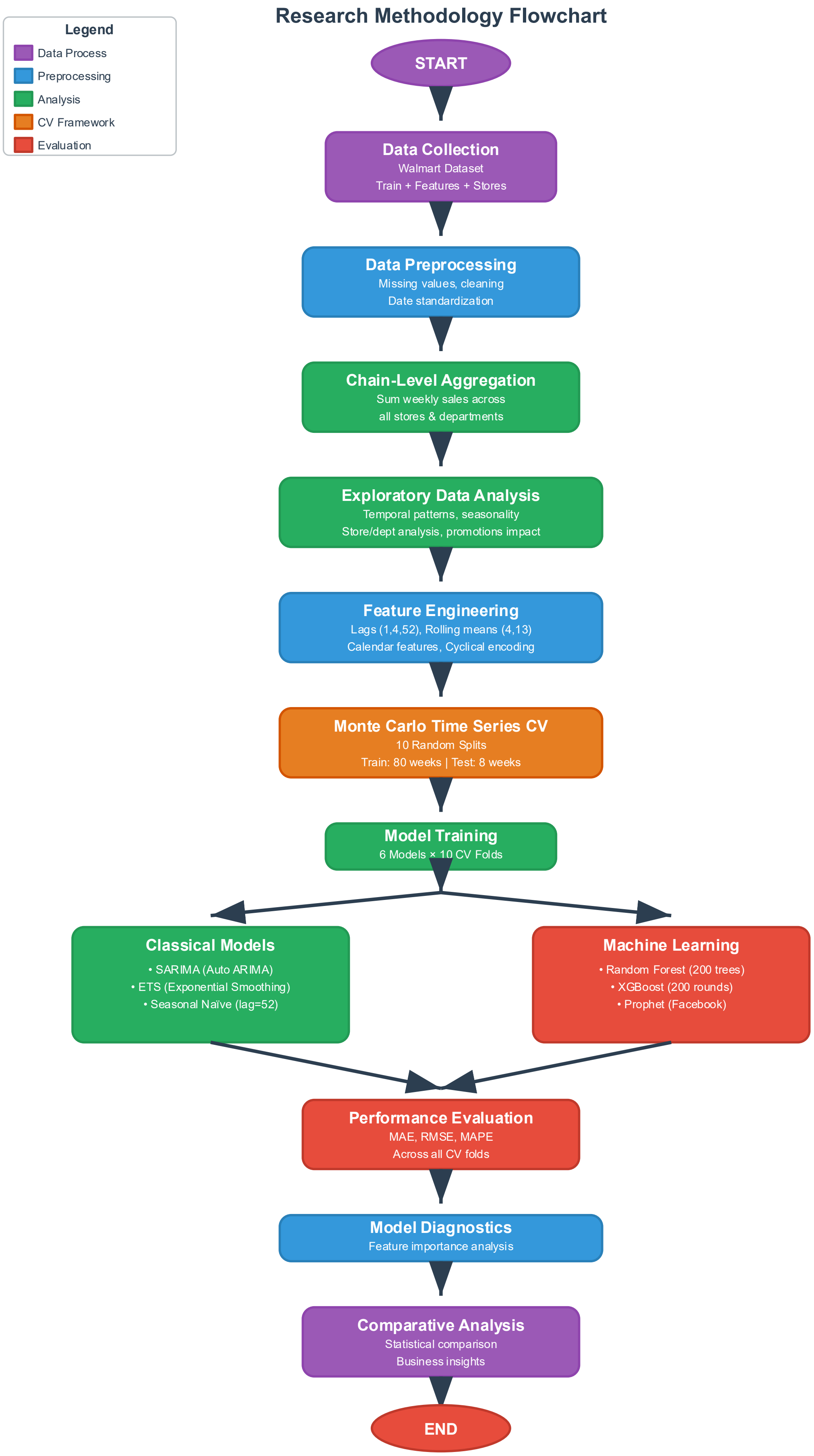
Research Methodology Flowchart
Legend
🟪 Data Process
🟦 Preprocessing
🟩 Analysis
🟧 CV Framework
🟥 Evaluation
START
Data Collection
Walmart Dataset
Train + Features + Stores
Data Preprocessing
Missing values, cleaning
Date standardization
Chain-Level Aggregation
Sum weekly sales across
all stores & departments
Exploratory Data Analysis
Temporal patterns, seasonality
Store/dept analysis, promotions impact
Feature Engineering
Lags (1,4,52), Rolling means (4,13)
Calendar features, Cyclical encoding
Monte Carlo Time Series CV
10 Random Splits
Train: 80 weeks | Test: 8 weeks
Model Training
6 Models × 10 CV Folds
Classical Models
• SARIMA (Auto ARIMA)
• ETS (Exponential Smoothing)
• Seasonal Naïve (lag=52)
Machine Learning
• Random Forest (200 trees)
• XGBoost (200 rounds)
• Prophet (Facebook)
Performance Evaluation
MAE, RMSE, MAPE
Across all CV folds
Model Diagnostics
Feature importance analysis
Comparative Analysis
Statistical comparison
Business insights
END

Figure 1: Methodology Flow Chart

3.2 Data

The analysis integrates three primary datasets from Walmart's retail operations spanning February 2010 through October 2012:

Primary Data Sources:

- Training Dataset (train.csv): 421,570 records containing Store, Department, Date, Weekly_Sales, and IsHoliday variables
- Features Dataset (features.csv): 8,190 records with external factors including Temperature, Fuel_Price, MarkDown1-5, CPI, Unemployment, and IsHoliday indicators
- Stores Dataset (stores.csv): 45 records providing Store, Type (A/B/C classification), and Size metadata

The dataset represents 143 weeks of observations across 45 stores and 81 departments, providing authentic retail complexity through multiple store formats, extensive promotional activities, and rich external economic variables.

Each store has multiple departments, but for this study we aggregated sales to the **chain level**, i.e. summing Weekly_Sales across all stores to get a single time series of total company-wide sales per week. This aggregate series smooths out some noise and highlights overall seasonal trends, providing a challenging but stable target for forecasting. We focus on chain-level demand since that is crucial for high-level planning (and was a key output in the M5 competition results)

In addition to sales, the dataset provides several exogenous variables by week and store (Walmart's *"features"* dataset):

Feature Name	**Data Type**	**Description**
Holiday indicator	Boolean	Flag indicating if the week contained a major holiday affecting sales (Super Bowl, Labor Day, Thanksgiving, Christmas). For chain-level data, treated as holiday week if any store had the holiday flag
Temperature	Numeric (Fahrenheit)	Average weekly temperature in the region of each store, capturing weather-related demand variations
Fuel Price	Numeric (USD)	Price of gasoline in the local region serving as an important economic indicator that might influence shopping behavior and disposable income

Feature Name	Data Type	Description
CPI (Consumer Price Index)	Numeric (Index)	Regional inflation indicator reflecting cost of typical goods, provided per region to capture local economic conditions
Unemployment rate	Numeric (Percentage)	Local unemployment percentage given per store region, representing economic health and consumer spending capacity
Markdowns	Numeric (5 variables, USD)	Five markdown variables indicating promotional discount events and their dollar magnitude at stores, primarily corresponding to holiday promotions. Many values are missing (NaN) for weeks without promotions

Table 1: Walmart Store Features: Exogenous Variables for Sales Forecasting

3.2.1 Data Merging Process

The data integration process was implemented through sequential left joins to preserve the complete temporal structure of sales observations, with Store and Date serving as primary keys for joining operations.

The merging process combined sales data with corresponding temporal features and store characteristics, creating a comprehensive dataset for analysis. During the initial data exploration, duplicate holiday indicators were identified and resolved by retaining the IsHoliday variable from the training dataset.

3.2.2 Comprehensive Missing Value Analysis

Variable	Missing Count	Missing Percentage
MarkDown2	310,322	73.61%
MarkDown4	286,603	67.98%
MarkDown3	284,479	67.48%
MarkDown1	270,889	64.26%
MarkDown5	270,138	64.08%

Table 2: Missing value table

A comprehensive missing data analysis revealed that promotional markdown variables (MarkDown1-5) contained substantial missing values, with MarkDown2 showing the highest missingness at 73.6%, followed by MarkDown4 (68.0%), MarkDown3 (67.5%), MarkDown1 (64.3%), and MarkDown5 (64.1%).

All other variables in the merged dataset showed complete data coverage with no missing values.

The missing values in markdown variables were strategically imputed with zeros, based on the business logic that missing markdown data indicates the absence of promotional activities during those periods. This approach aligns with retail practice where promotional markdowns are discrete events rather than continuous activities. The imputation strategy was validated by examining the temporal distribution of missing values, which confirmed that markdown data was only available from November 2011 onwards, consistent with the dataset documentation.

Following imputation, the dataset achieved complete coverage with no remaining missing values across all 421,570 observations and 16 variables.

3.2.3 Temporal Feature Engineering

To support both classical time series models and machine learning approaches, comprehensive temporal features were engineered from the Date variable. These features capture multiple levels of seasonality and temporal patterns essential for retail sales forecasting.

Temporal Identifier	Description
Year	**Annual trends and year-over-year growth patterns**
Quarter	**Quarterly business cycles and seasonal patterns**
Week	**ISO week numbers for weekly seasonality analysis**
Month	**Monthly seasonal patterns with abbreviated labels**
Day of week	**Weekly cyclical patterns in retail sales**
Year-month	**Combined temporal identifier for monthly aggregations**

Table 3: Temporal Feature Engineering

3.2.4 Negative Sales Investigation

A detailed exploratory data analysis (EDA) on the Walmart sales dataset revealed 1,285 instances of negative weekly sales, constituting 0.305% of the total observations. These negative values are not errors or data corruption, but represent legitimate business events such as product returns, refunds, and inventory adjustments.

Statistical Summary of Negative Sales

- Total Negative Observations: 1,285 (0.305% of dataset)
- Value Range: –$4,988.94 to –$0.02
- Mean Negative Sales: –$68.61
- Median Negative Sales: –$13.20

The median being substantially smaller in magnitude than the mean indicates that most negative sales represent small returns, while a few larger returns skew the average downward, consistent with typical retail return behavior.

Temporal Patterns

Analysis of negative sales by year and month revealed distinct temporal clusters aligned with business cycles:

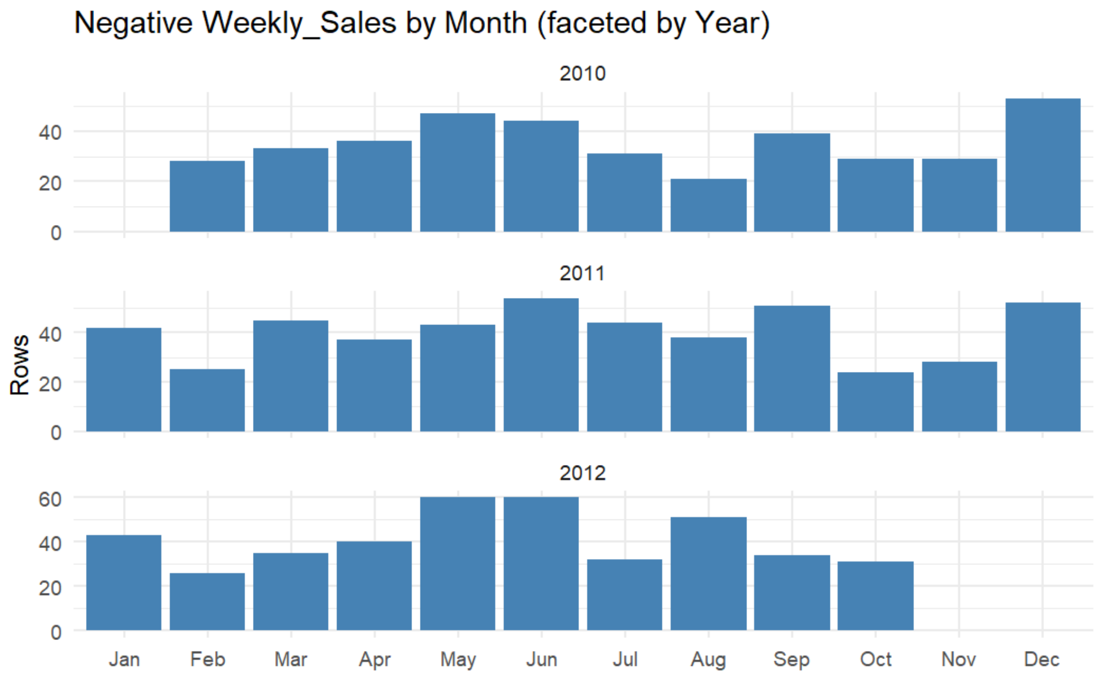

Figure 2: Negative weekly_sales by month

Peak Period	**Negative Sales Count**	**Business Explanation**
May 2011	60	Mid-year clearance / seasonal returns
June 2011	60	Mid-year clearance / seasonal returns
December 2011	52	End-of-year adjustments
December 2010	53	Inventory corrections pre-holiday

Table 4: Peak Negative Sales Periods

Outside of peak periods, negative sales averaged between 20 to 40 observations monthly, reflecting normal return activity.

Store-Level Distribution

Negative sales were proportionally distributed across store types, indicating systemic operational characteristics rather than anomalies:

Store Type	**Total Records**	**Negative Count**	**Negative Rate**
Type B	163,495	676	0.41%
Type C	42,597	124	0.29%
Type A	215,478	485	0.23%

Table 5: Negative Sales Distribution by Store Type

Type B stores exhibited slightly higher return rates, potentially attributable to product mix or customer demographics.

Departmental Analysis

Several departments showed a concentration of negative sales, aligning with industry expectations regarding return likelihood:

Department	**Negative Count**	**Likely Category**	**Return Logic**
Dept 47	**254**	**Electronics/Seasonal**	**Technical issues/returns**
Dept 18	**180**	**Consumer goods**	**Defective/unwanted items**
Dept 54	**146**	**Apparel/Clothing**	**Size/fit issues**
Dept 19	**87**	**Seasonal merchandise**	**Weather-dependent returns**

Table 6: Negative Sales by Department

The department-wise concentration validates that returns are related to product characteristics rather than data irregularities.

Validation That Negative Sales Are Legitimate Business Events

1. Systematic Temporal Patterns:
   Errors would be randomly distributed over time; however, negative sales cluster in predictable seasonal periods such as holiday and clearance seasons.
2. Reasonable Value Ranges:
   Negative values remain within plausible business limits, with the largest negative sale recorded at –$4,988.94, consistent with bulk returns rather than erroneous extremes.
3. Departmental Concentration:
   Returns predominantly affect specific departments known for higher return rates, contradicting the random distribution expected from data errors.
4. Proportional Distribution Across Store Types:
   Negative sales occur consistently across all store types, indicating uniform business processes rather than isolated anomalies.

What Weekly_Sales Represents

The Weekly_Sales variable effectively captures net revenue, calculated as gross sales minus returns, refunds, inventory adjustments, and vendor credits. For example, a negative weekly sale in Department 47, Store 3 for the week of January 8, 2011 (–$2,500) likely represents significant returns exceeding new sales, consistent with post-holiday customer behavior.

Handling Negative Sales in Analysis

Rather than excluding these legitimate negative sales—which would distort revenue patterns and underrepresent return activities—this study retains these entries to preserve the true shape of sales variability. Robust modeling techniques capable of accommodating such business realities, including seasonal return patterns and outlier handling, are employed to ensure analytical accuracy.

3.3 Exploratory Data Analysis Findings

3.3.1 Sales Patterns

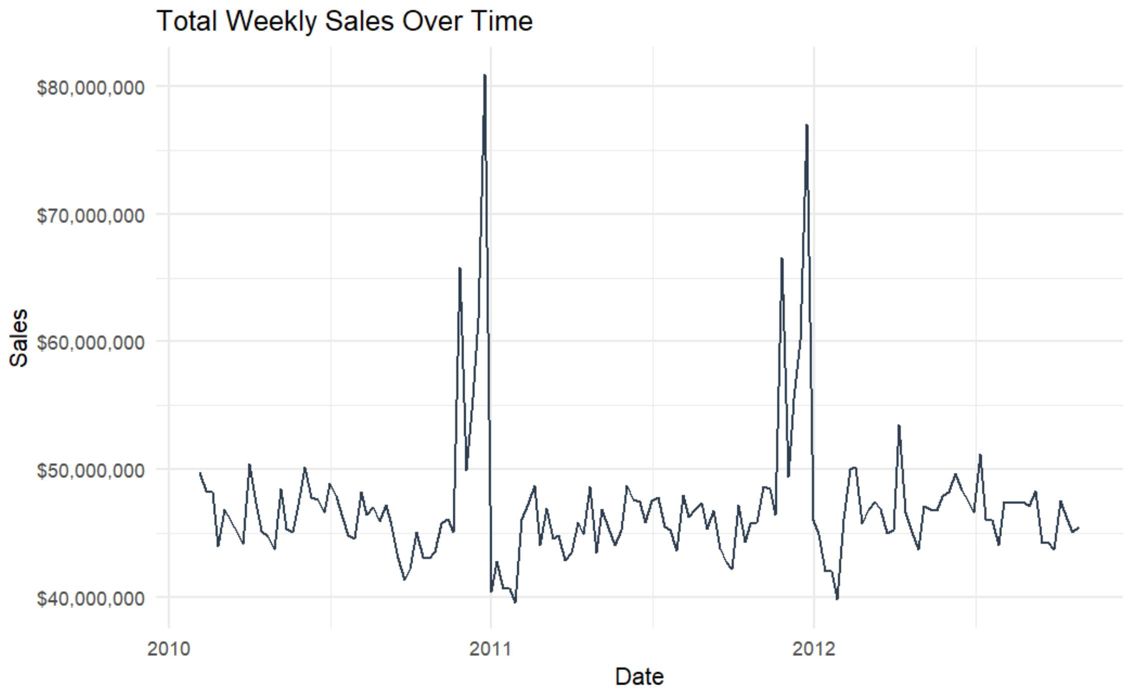

Figure 3: Total weekly sales over time

The **chain-level weekly series** (Figure: *Total Weekly Sales Over Time*) shows a relatively stable band through spring/summer with two dramatic spikes per fiscal year in late November and late December (the retail holiday season). During regular periods throughout the year, the company maintained steady baseline sales ranging between $40-50 million per week, demonstrating consistent operational performance. However, this stability was dramatically punctuated by explosive holiday spikes that reached $80+ million during Thanksgiving and Christmas weeks, representing a remarkable 100% increase over normal sales levels. Those spikes are consistent across FY-2010 and FY-2011; FY-2012 terminates in late October due to the competition's test horizon being withheld. The level of the band varies modestly year-to-year, motivating "year" as a low-frequency control in ML models and suggesting small shifts in intercept or local trend components for classical models. Following each holiday peak, sales rapidly normalized back to baseline levels, suggesting that the increased purchasing was primarily driven by seasonal demand rather than sustained growth. The summer months showed relative stability with only moderate variability, providing a period of predictable revenue flow.

3.3.2 Average weekly sales by store type

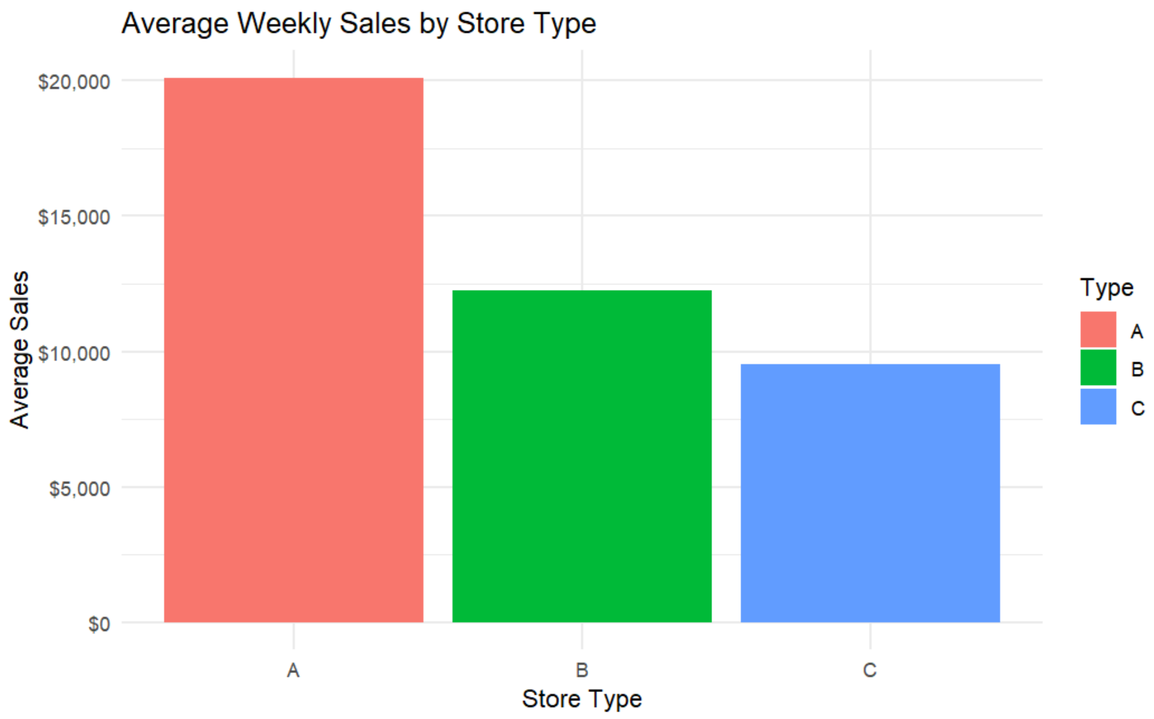

Figure 4: Average weekly sales by store type

Store Type	Average Weekly Sales	Performance Tier
Type A	$20,000	Premium/Supercenters
Type B	$12,000	Mid-tier/Regional
Type C	$9,500	Neighborhood/Local

Table 7: Average weekly sales by store type

The store performance analysis reveals a strategic three-tier retail hierarchy that reflects deliberate market positioning and operational scaling. Type A stores, functioning as premium supercenters or flagship locations, dominate the performance landscape with average weekly sales of $20,000, establishing them as the company's revenue powerhouses. Mid-tier Type B stores, likely serving regional markets, generate $12,000 in average weekly sales, positioning them as solid performers that bridge the gap between premium and local operations. Type C stores, representing neighborhood or local market presence, contribute $9,500 in weekly sales, providing essential community-level market penetration. This performance hierarchy creates a significant 2.1× sales ratio between the highest and lowest performing store types, demonstrating the substantial impact of store format and market positioning on revenue generation. This structured approach to store performance creates both opportunities for targeted investment in high-performing formats and challenges for inventory management and demand forecasting across diverse retail environments.

3.3.3 Variable Correlation Structure Analysis

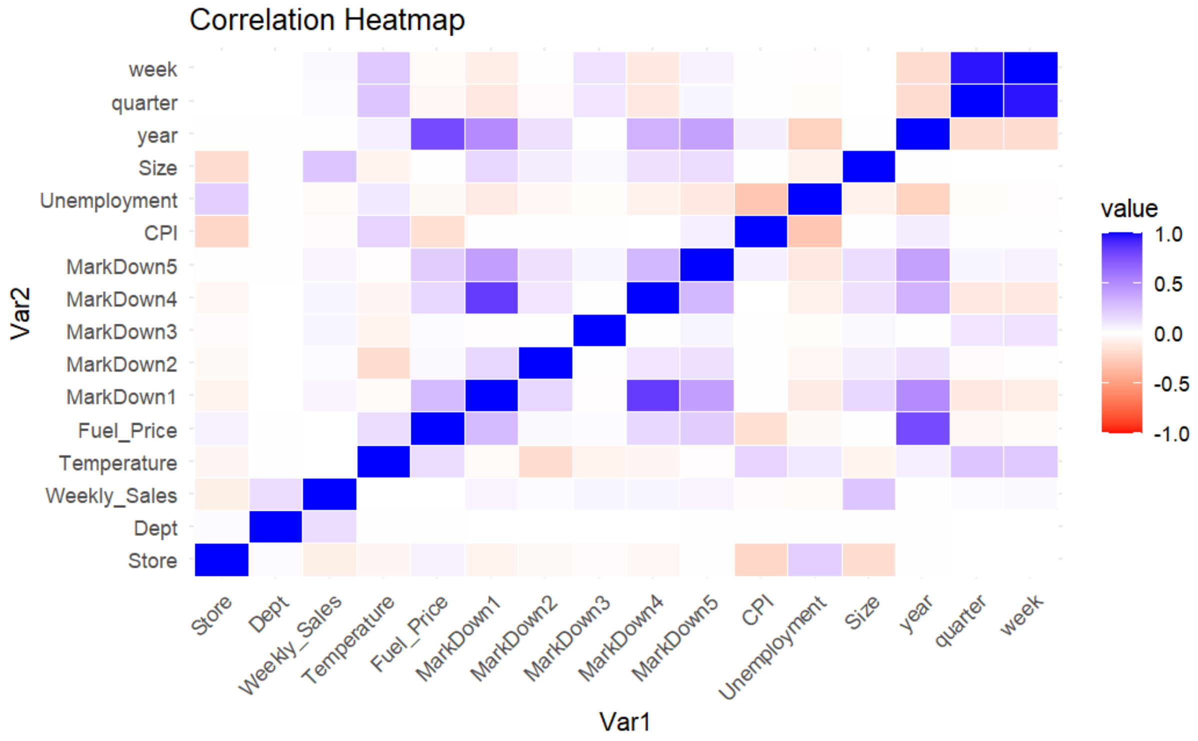

Figure 5: Correlation matrix

The correlation analysis reveals sophisticated relationships within the retail data ecosystem. Temporal variables create the strongest patterns, with quarter and week displaying exceptional correlation (0.964), reflecting their mathematical interdependence in capturing seasonal retail patterns.

Promotional variables show remarkable internal coordination: MarkDown1 and MarkDown4 exhibit strong correlation (0.839), indicating synchronized discount strategies across product categories. Economic relationships emerge through fuel price and year correlation (0.780), confirming consistent inflationary pressures. Several markdown variables demonstrate moderate positive correlations with year—MarkDown1 (0.501), MarkDown5 (0.403), and MarkDown4 (0.335)—suggesting promotional activities have intensified over time.

Most significantly, weekly sales shows surprisingly weak correlations across all variables, with strongest relationships being modest connections to store size (0.244) and department (0.148). The absence of strong correlations between sales and markdown variables indicates promotional pricing strategies don't translate into immediate linear sales improvements. The negative CPI-unemployment correlation (-0.300) reflects expected macroeconomic dynamics. These patterns suggest retail success depends on complex, potentially non-linear interactions between variables rather than simple linear relationships.

3.3.4 Sales Distribution Characteristics

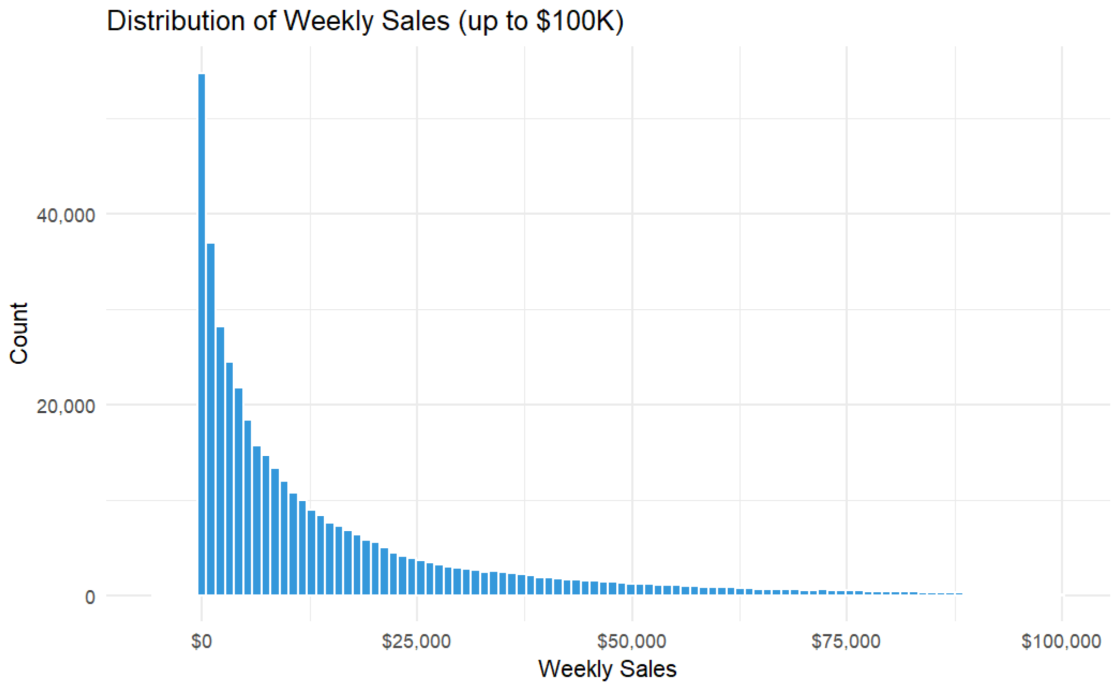

Figure 6: Distribution of weekly sales

Weekly sales distribution reveals a classic retail pattern: heavily right-skewed with extended tail demonstrating fundamental performance inequality across the network. The distribution peaks at near-zero sales with approximately 45,000+ observations clustered in minimal sales ranges, reflecting that many sales records across stores, periods, or categories generate modest weekly volumes.The distribution exhibits dramatic exponential decay from the initial peak to progressively smaller frequencies as sales increase. The long tail captures exceptional high-performance periods—seasonal peaks, promotional events, or high-performing locations—that generate disproportionate sales volumes dramatically impacting overall metrics.

This pattern underscores forecasting challenges: the vast majority of observations follow predictable low-volume patterns while a critical minority of exceptional periods generate revenue spikes determining business success. This creates complex dynamics between consistent baseline performance and intermittent high-impact events, highlighting operational realities where performance varies due to seasonal fluctuations, inventory constraints, and natural business cycles.

3.3.5 weekly sales spikes by store type

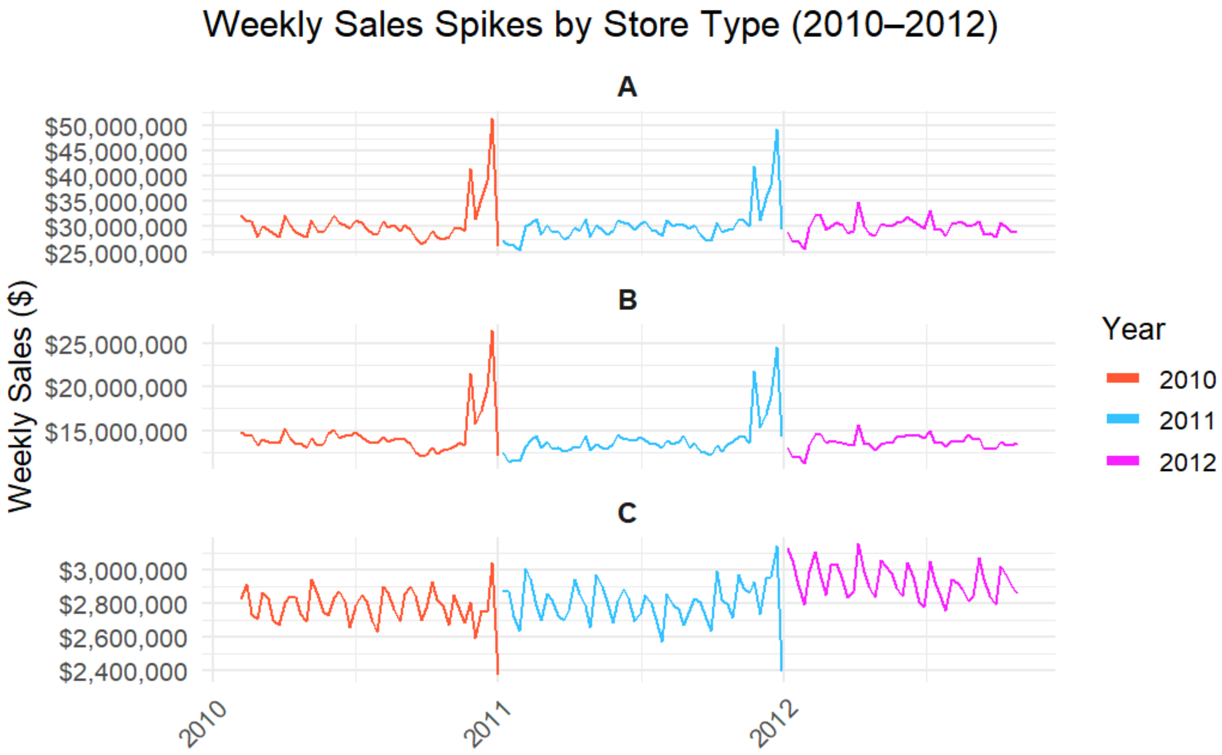

Figure 7: weekly sales spikes by store type

The weekly sales spikes by store type analysis revealed distinct performance hierarchies and seasonal response patterns across Walmart's retail formats.

Type A Supercenters dominated with $25-50 million weekly sales, exhibiting the most pronounced seasonal spikes (reaching $50 million during holidays) and highest promotional sensitivity, indicating price-responsive customer behavior.

Type B Discount Stores occupied the middle tier at $12-25 million weekly, showing similar but proportionally smaller seasonal patterns with more stable baselines and consistent holiday lift across years.

Type C Neighborhood Markets demonstrated remarkable stability at $2.8-3.2 million weekly with minimal seasonal variation and limited promotional response, reflecting convenience-focused rather than seasonal shopping behaviors.

This analysis confirms store type as a fundamental segmentation variable transcending simple size differences. Each format exhibits distinct seasonal sensitivity profiles reflecting unique market positioning: supercenters capture high-volume seasonal shoppers, discount stores serve price-conscious but seasonally-aware customers, and neighborhood markets provide steady convenience-driven revenue streams, creating a strategic portfolio optimizing market coverage and revenue stability.

3.4 Seasonality and Cyclical Pattern Analysis

3.4.1 Fiscal Year Seasonality

Walmart’s fiscal year (FY) runs from 1 February to 31 January, and FY labels refer to the calendar year in which the fiscal year ends (e.g., FY2012 spans 1 Feb 2011–31 Jan 2012). Quarters are aligned to the retail calendar as: Q1 = Feb–Apr, Q2 = May–Jul, Q3 = Aug–Oct, Q4 = Nov–Jan. **Alignment with the Kaggle Walmart dataset.** The Kaggle “Walmart Store Sales” data are **weekly** and date-stamped on **Fridays** (week-ending). Your analysis window (2010-02-05 to 2012-10-26) therefore spans portions of **FY2011–FY2013** under Walmart’s fiscal convention:

- 2010-02-05 falls in **FY2011 Q1**,
- all weeks in Nov–Jan fall in **Q4** of the corresponding fiscal year (e.g., Black Friday and Christmas are in **FY2012 Q4** for late-2011 weeks),
- weeks in Aug–Oct map to **Q3**, which often captures “back-to-school” demand.

This fiscal framing is important for two reasons. First, **seasonality in retail is fiscal-calendar driven**, not purely Gregorian: Thanksgiving, Black Friday, Christmas, and back-to-school sit in fixed **fiscal** quarters, and promotion calendars follow the 4-5-4 cadence. Second, **year-over-year (YoY)** comparisons in retail practice compare **fiscal week t** this year to **fiscal week t** last year (weekday-aligned), not simply the same date range.

3.4.2 Fiscal Year Seasonality Assessment

Q. Does total sales differ between months, Are there seasonal differences by time period?

a. Monthly Sales Patterns:

For monthly seasonality analysis, I used Feb 2010 - Jan 2012 (2 complete fiscal years): Every month has exactly 2 years of data for fair comparison, Captures complete seasonal cycles in Walmart's fiscal calendar, Equal representation prevents bias toward certain month

Fiscal Month Performance Results:

Fiscal Month	Sales ($M)	Business Period
FM 1 (Feb)	376.7	Post-holiday recovery
FM 2 (Mar)	361.3	Spring baseline
FM 3 (Apr)	457.9	Spring peak
FM 10 (Nov)	413.0	Pre-holiday buildup
FM 11 (Dec)	576.8	Holiday peak

Fiscal Month	Sales ($M)	Business Period
FM 12 (Jan)	332.6	Post-holiday low

Table 8: Fiscal Month Performance Results

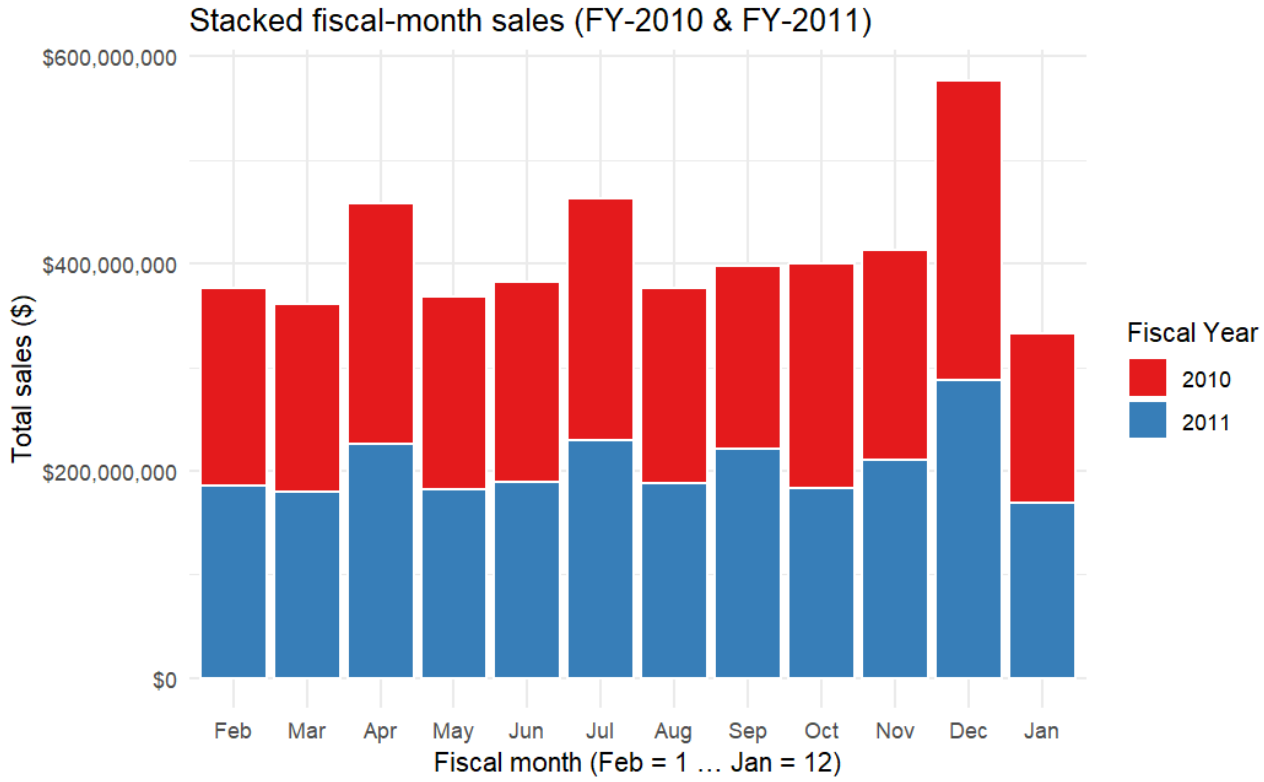

Figure 8: Stacked Fiscal Month Sales

The fiscal month sales analysis across two complete fiscal years (FY-2010 and FY-2011) reveals distinct seasonal patterns that provide crucial insights into Walmart's revenue cycles and customer behavior dynamics. The stacked bar chart demonstrates clear monthly variations in total sales, with December emerging as the undisputed peak month, generating $576.8 million across the two-year period, representing a dramatic 73% increase over the lowest-performing month of March ($361.3 million). This December surge reflects the powerful impact of holiday shopping, Black Friday promotions, and year-end consumer spending that fundamentally drives retail performance.

Fiscal Month	Month Name	Total Sales ($)	% of Average
11 (Peak)	December	$576,838,635	173%
12 (Trough)	January	$332,598,438	100%
Gap	**Peak-Trough**	**$244,240,197**	**73% increase**

Table 9: Monthly Seasonality (Fiscal Year Basis)

The seasonal narrative unfolds through three distinct phases: a winter decline period from February through May where sales gradually decrease from $376.7 million to $368.4 million, suggesting post-holiday recovery and spring shopping moderation; a summer stabilization period from June through August where sales remain relatively consistent between $382-462 million, indicating steady but unremarkable consumer activity; and an autumn acceleration period from September through December where sales progressively climb from $398.1 million to the peak December performance. November represents a critical transition month with $413.0 million in sales, serving as the launchpad for the holiday surge while October ($400.4 million) and September ($398.1 million) show the beginning of seasonal momentum building. The year-over-year comparison within each month, visible through the blue (2011) and red (2010) stacked segments, reveals generally consistent seasonal patterns across both fiscal years, though with some notable variations that suggest evolving market conditions or strategic initiatives. January shows interesting dynamics as the fiscal year-end month with $332.6 million, representing the post-holiday normalization that sets the stage for the subsequent February recovery. Similar patterns across FY-2010 and FY-2011

b. Weekly Seasonal Patterns

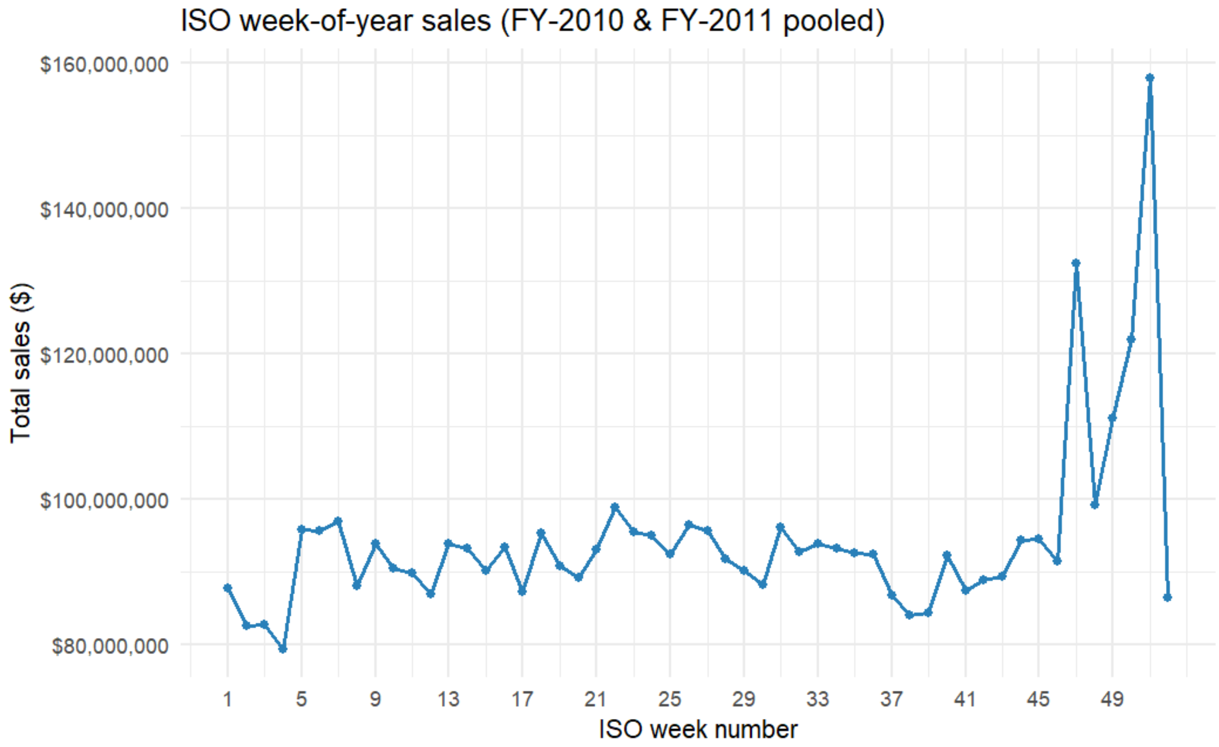

Figure 9: Weekly Seasonal patterns FY 2010-11 pooled

ISO week analysis across pooled FY-2010 and FY-2011 data reveals both baseline stability and extraordinary seasonal volatility throughout Walmart's fiscal calendar. For weeks 1-45 (87% of the year), total sales maintain stable baseline ranging $79-99 million per week, averaging $90-95 million with occasional peaks reaching $99 million during weeks 22-23 (late spring/early summer).

The most striking pattern emerges in the final seven weeks (46-52), where sales undergo dramatic transformation. Week 47 marks the holiday surge beginning at $132.4 million (33% increase over baseline), followed by progressive growth: week 48 ($99.3M), week 49 ($111.2M), week 50 ($121.9M), and week 51's extraordinary peak of $157.9 million—a remarkable 75% increase over baseline levels. Week 52 concludes at $86.5 million, showing immediate post-holiday normalization.

This pattern reveals that while Walmart maintains consistent operational performance for 87% of the fiscal year, the final 13% (holiday season) generates disproportionate revenue impact representing 20-30% above normal weekly volumes.

Weekly Amplitude
The seasonal amplitude at the weekly level is even more pronounced than monthly patterns:

- Peak-to-trough variation: Week 51 ($157.9M) versus Week 4 ($79.4M) represents a 99% increase from trough to peak
- This extreme variation poses significant challenges for forecasting models, particularly during transition periods

3.4.3 Year-over-Year Performance Analysis

Q.Are there statistically significant differences in sales performance across fiscal years 2010-2012?
The fiscal year comparison analysis examined sales performance consistency across the observation period, focusing on comparable months (February through October) to ensure valid year-over-year comparisons.

Annual Performance Comparison:

Fiscal Year	Mean Weekly Sales ($)	Standard Deviation ($)	Sample Size
FY2010	**$46,083,903**	**$2,223,142**	**39 weeks**
FY2011	**$45,801,425**	**$1,762,657**	**39 weeks**
FY2012	**$46,954,830**	**$2,056,149**	**39 weeks**

Table 10: Annual performance comparison

Year-over-Year Changes:

- FY2010 → FY2011: -$282,478 (-0.6%)
- FY2011 → FY2012: +$1,153,405 (+2.5%)

Percentage changes were computed year-over-year to quantify the direction and magnitude of shifts, Overall trend: Modest recovery after slight decline.

Statistical Test Results

- Test Used: One-way ANOVA (normality assumption met)
- F-statistic: 3.445
- p-value: 0.035 (statistically significant at $\alpha = 0.05$)
- Effect Size: Partial $\eta^2 = 0.057$ (small to medium effect)

The fiscal year-over-year analysis reveals statistically significant but practically modest differences in Walmart's sales performance across the three-year period, with the one-way ANOVA indicating meaningful variation between fiscal years (F(2,23) = 3.445, p = 0.035). However, the effect size remains small (partial $\eta^2 = 0.057$), indicating that only 5.7% of the variance in monthly sales can be attributed to fiscal year differences, while the remaining 94.3% stems from other factors including seasonal patterns, economic conditions, and operational variables.

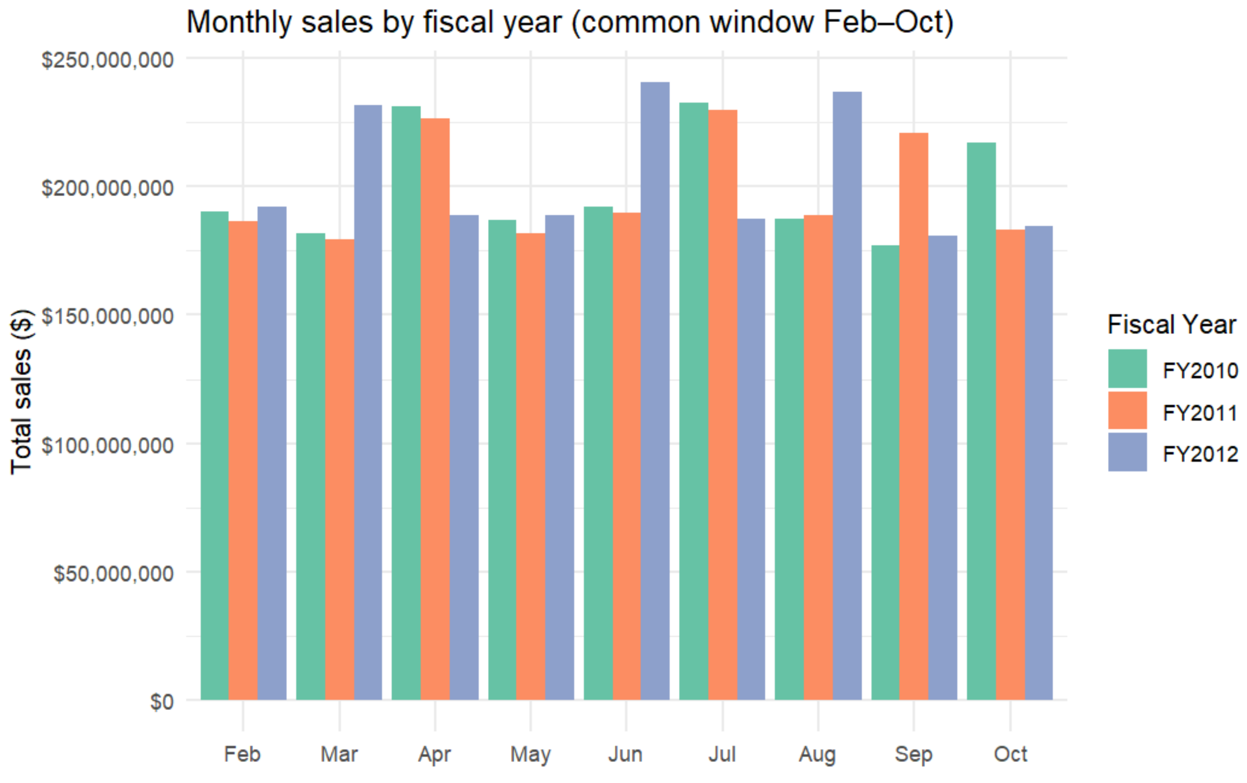

Figure 10: Month-by-Month Sales Comparison

Month	FY2010 ($)	FY2011 ($)	FY2012 ($)	Highest Year	Variation (%)
Feb	190,332,983	186,331,328	192,063,580	FY2012	3.1%
Mar	181,919,803	179,356,448	231,509,650	FY2012	29.1%
Apr	231,412,368	226,526,511	188,920,906	FY2010	22.5%
May	186,710,934	181,648,158	188,766,479	FY2012	3.9%
Jun	192,246,172	189,773,385	240,610,329	FY2012	26.8%
Jul	232,580,126	229,911,399	187,509,452	FY2010	24.0%
Aug	187,640,111	188,599,332	236,850,766	FY2012	26.2%
Sep	177,267,896	220,847,738	180,645,544	FY2011	24.6%
Oct	217,161,824	183,261,283	184,361,680	FY2010	18.5%

Table 11: Fiscal Year Monthly Sales Totals with Cross-Year Variation

The month-by-month variation analysis exposes dramatic inconsistencies in seasonal performance patterns that drive the statistical significance. March emerges as the most volatile month with a staggering 29.1% difference between years, ranging from FY2011's low of $179.4 million to FY2012's exceptional peak of $231.5 million, representing a $52 million swing that suggests either extraordinary promotional success or fundamental market disruption. Conversely, February demonstrates remarkable stability with only 3.1% variation across years ($186.3-192.1 million), while May shows similar consistency at 3.9% variation, indicating these months may be less susceptible to external market forces or strategic initiatives.

FY2012 displays particularly erratic behavior, claiming the highest performance in five of nine months (March, May, June, August, and February) while simultaneously recording the lowest performance in April and July, creating an unprecedented pattern of monthly volatility. This irregular seasonal behavior contrasts sharply with FY2010's more traditional retail pattern of strong April, July, and October performance, and FY2011's relatively stable month-to-month consistency. The year-month interaction analysis reveals poor seasonal consistency across the dataset, with different fiscal years peaking in entirely different months, suggesting that external economic factors, competitive pressures, or internal strategic changes fundamentally altered Walmart's traditional seasonal sales patterns during this period.

The statistical significance (p = 0.035) confirms that mean monthly sales differ meaningfully across the three fiscal years when restricted to the February-October window, though the small effect size indicates these differences, while statistically detectable, represent relatively modest practical impact on overall business performance. The analysis demonstrates that while Walmart maintained consistent average monthly performance levels ($201.7-207.0 million), the underlying monthly patterns shifted dramatically between years, creating forecasting challenges and indicating that traditional seasonal models would require year-specific adjustments to maintain accuracy during this volatile period.

3.4.4 Store-Level Performance Analysis

Q.Which stores are performing best/worst and How do stores compare systematically?

Store Performance Rankings:

Category	**Store**	**Total Sales ($)**	**Rank Overall**	**Rank FY 2012**	**Performance Multiple**
Top 3	Store 20	$301,397,792	1	2	8.1× bottom performer
	Store 4	$299,543,953	2	1	8.1× bottom performer
	Store 14	$288,999,911	3	6	7.8× bottom performer
Bottom 3	Store 5	$45,475,689	43	42	-$256M from top
	Store 44	$43,293,088	44	43	-$258M from top
	Store 33	$37,160,222	45	45	-$264M from top

Table 12: Store performance

Most Consistent Top Performer: Store 20 (ranks 1-2-1)

Most Consistent Bottom Performer: Store 33 (rank 45 all years)

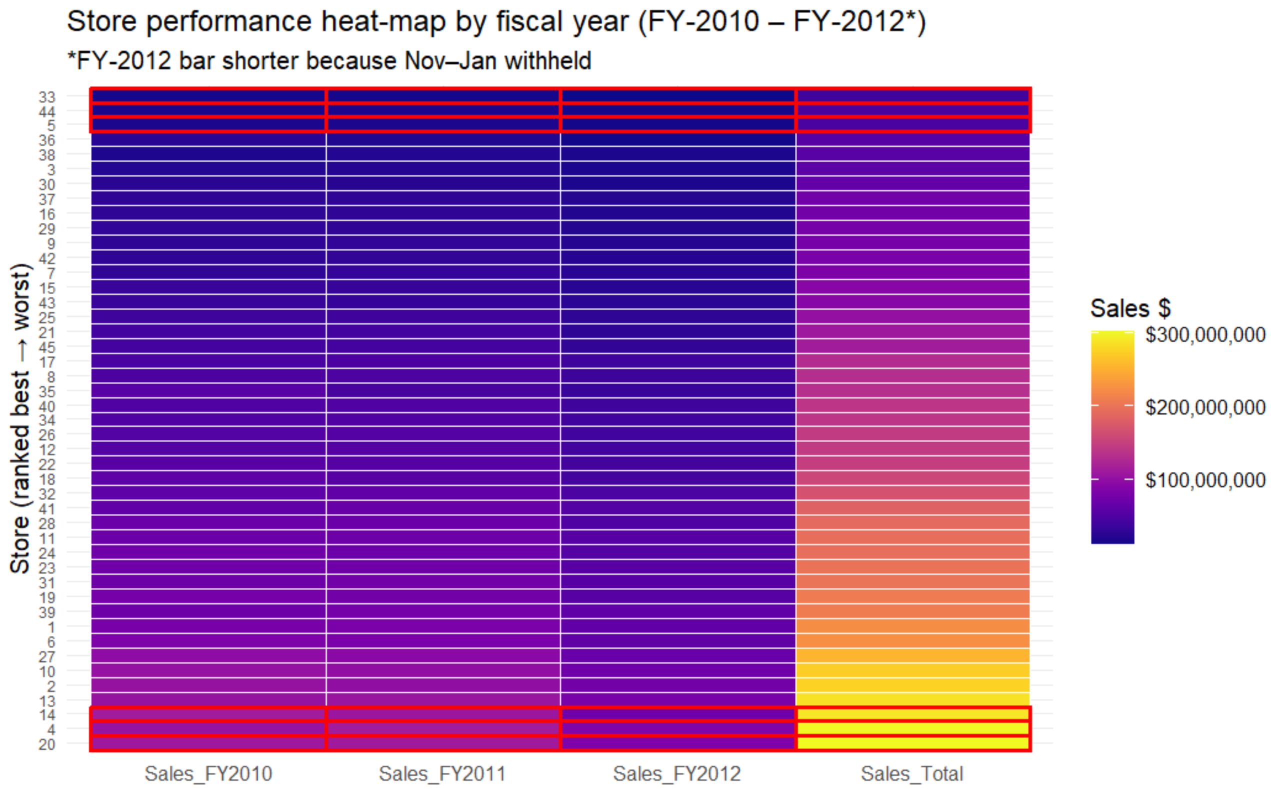

Figure 11: Store performance heat-map

The store performance heatmap reveals a pronounced and persistent hierarchy demonstrating remarkable consistency across fiscal years, with clear stratification between top-performing, middle-tier, and underperforming stores.

Store-level analysis exposes dramatic performance disparities within Walmart's retail network. The top performer (Store 20) generated $301.4 million in total sales compared to the bottom performer (Store 33) with $37.2 million—an extraordinary 8:1 performance ratio highlighting fundamental differences in market positioning, location quality, or operational efficiency. Top-tier stores (Stores 20, 4, 14) consistently maintain sales exceeding $280 million, middle-tier performers cluster between $100-250 million, and bottom-tier stores struggle below $100 million.

The hierarchy appears remarkably stable across fiscal years, suggesting performance differentials stem from structural factors such as location demographics, market competition, or store format rather than temporary operational issues. While middle-tier stores show moderate variation, the overall hierarchy persists. Bottom performers consistently generated less than $50 million each, indicating potential candidates for strategic review. This stability suggests aggregated forecasting may be more reliable than individual store predictions.

3.4.5 Department Performance Analysis

Q. What is the sales distribution across departments and how stable is this distribution over time?

Department Performance Rankings:

Rank	Department	Total Sales ($)
1	Dept 92	$483,943,341
2	Dept 95	$449,320,162
3	Dept 38	$393,118,136
…	…	…
81	Dept 47	-$4,962.93

Table 13: Department performance rakings

Performance gap: $483,948,305

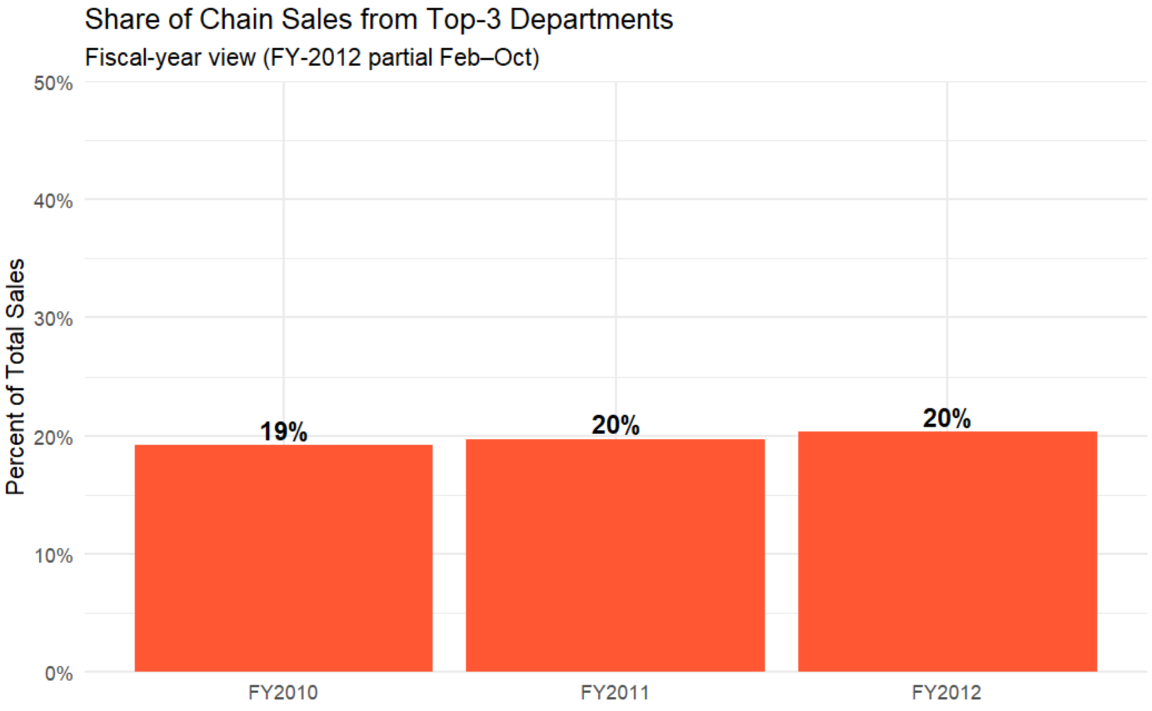

Figure 12: share of chain sales form top 3 departments

The department-level performance analysis reveals a highly concentrated sales distribution with extreme inequality that demonstrates remarkable stability across fiscal years, fundamentally shaping Walmart's revenue structure and strategic priorities. The top three departments (92, 95, and 38) maintain an extraordinary dominance, consistently generating approximately 19-20% of total chain sales each fiscal year, creating a combined 59-60% market share that represents the core revenue foundation of Walmart's operations. Department 92 emerges as the undisputed leader with $483.9 million in total sales across the observation period, followed closely by Department 95 ($449.3 million) and Department 38 ($393.1 million), establishing a clear performance hierarchy that persists despite the truncated FY2012 data.

Department	**Pooled Share of Rest**	**Mean Share of Rest (FY avg)**	**Pooled Share of TOTAL**
72	5.7%	5.6%	4.5%
90	5.4%	5.4%	4.3%
40	5.3%	5.4%	4.3%
2	5.2%	5.2%	4.2%
91	4.0%	4.0%	3.2%
13	3.6%	3.7%	2.9%
8	3.6%	3.6%	2.9%
94	3.5%	3.5%	2.8%
4	3.1%	3.1%	2.5%
93	3.0%	3.0%	2.4%
7	2.9%	2.8%	2.3%
79	2.6%	2.6%	2.1%
23	2.6%	2.6%	2.1%
5	2.5%	2.5%	2.0%
9	2.4%	2.4%	1.9%

Table 14: Revenue Distribution Among Non-Leading Departments (Top 15)

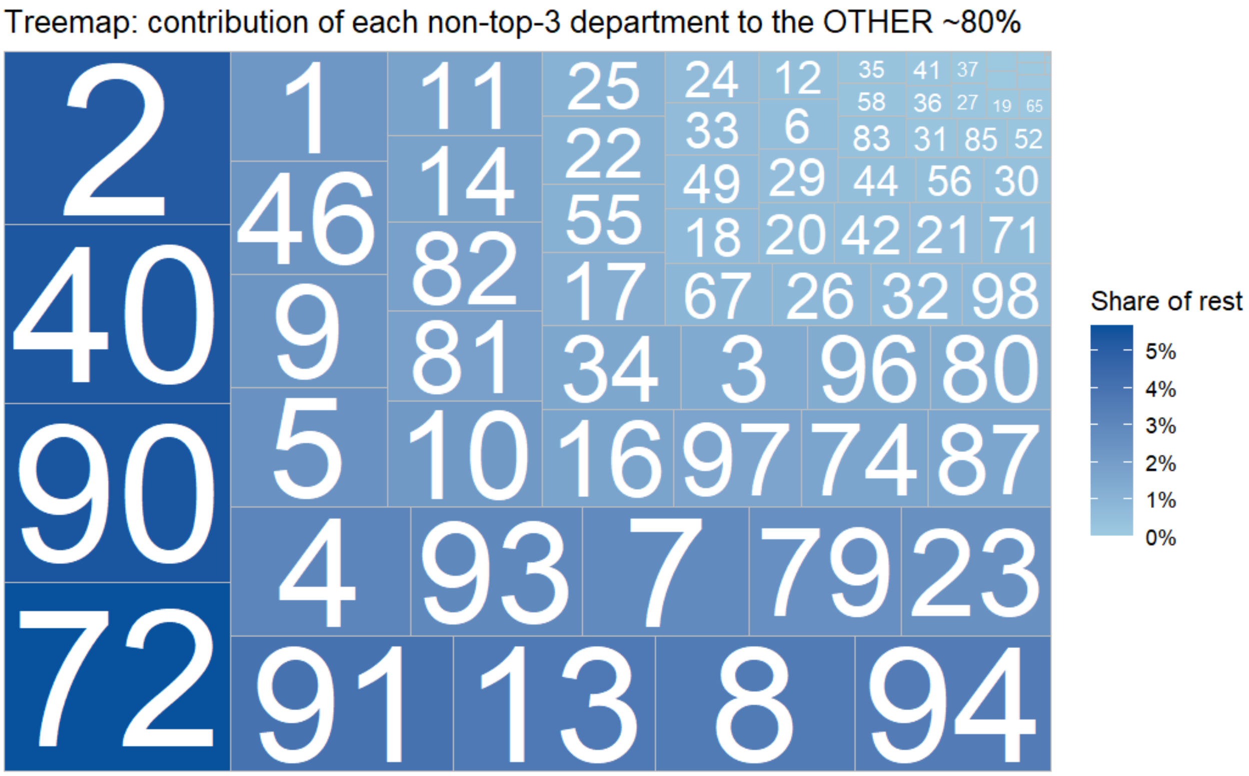

Figure 13: Treemap: Contribution of non-top 3 Departments

The treemap visualization provides an intuitive spatial representation of the departmentalhierarchy within the remaining 80% of sales, effectively illustrating the proportional contributions through rectangle sizes that correspond to each department's share of non-top-3 revenue. The visualization reveals a clear four-tier structure within the secondary departments: Department 72 occupies the largest rectangle with its 5.7% contribution of the non-top-three sales, representing 4.5% of total chain sales, establishing it as the clear leader among non-top-3 departments, followed by the substantial presence of Departments 90 (5.4%), 40 (5.3%), and 2 (5.2%) which form the second tier of major contributors. The third tier consists of moderately-sized departments including 91, 13, 8, and 94, each representing 3.0-4.0% of the remaining sales, while the fourth tier comprises smaller departments that collectively demonstrate the fragmented nature of the lower-performing segment.

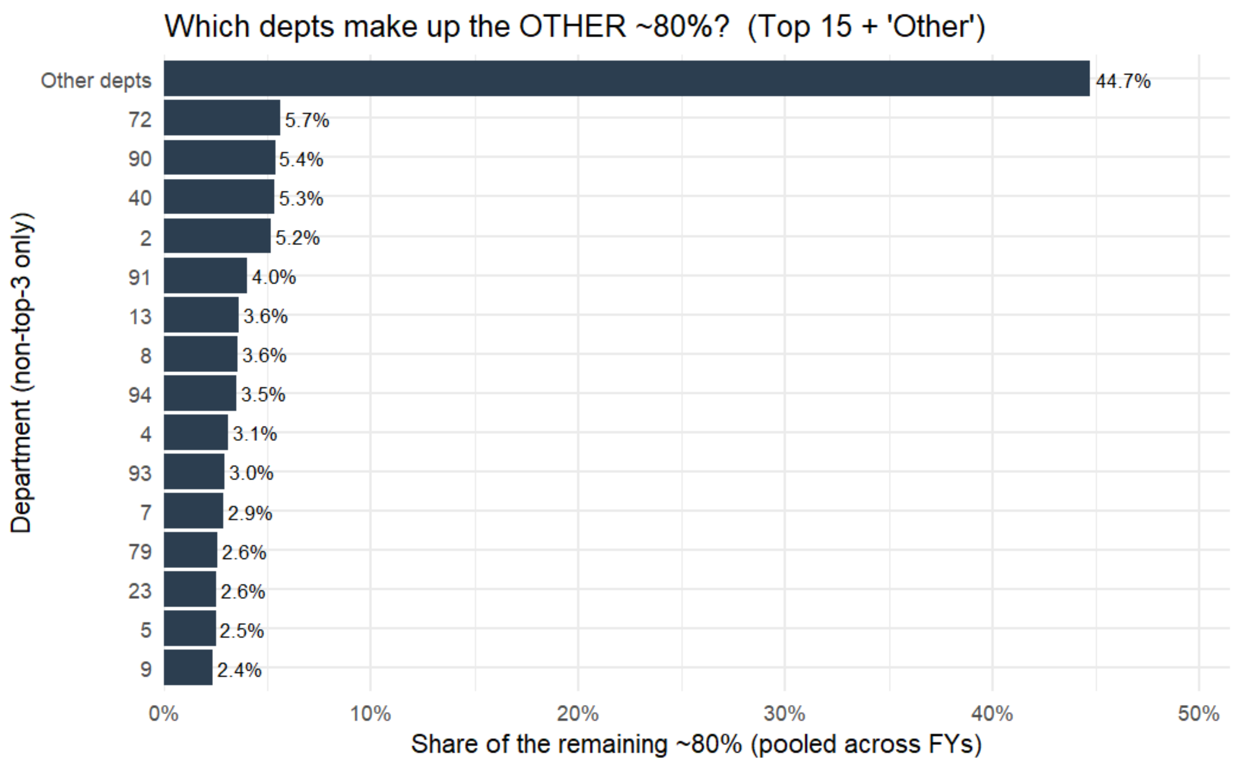

Figure 14: Secondary Departmental Revenue Distribution (Top 15 + Others)

The analysis reveals that 44.7% of the remaining sales comes from departments not included in the top 15 secondary performers, indicating that the majority of Walmart's departments operate as relatively small revenue contributors that collectively support the business but individually lack significant market impact.

This departmental concentration pattern exhibits remarkable temporal stability, with the top three departments maintaining nearly identical share percentages across all fiscal years despite varying market conditions and the partial FY2012 dataset.

This consistent ~20% concentration suggests that:

- Core departments maintain stable market positions over time
- Chain-level forecasting benefits from this stability, as major revenue drivers remain consistent
- Seasonal variations are likely driven by volume changes rather than fundamental shifts in department mix

3.4.6 Promotional and Holiday Impact Analysis

Weekly total sales were aggregated for the period February 5, 2010, through January 27, 2012, representing two complete fiscal years. Binary flags indicating the presence of promotions (based on any positive MarkDown indicator) and holiday weeks (based on IsHoliday) were created at the aggregate chain level for each week.

Non-Parametric Statistical Testing

Given the non-normal distribution and presence of outliers in retail sales data, Wilcoxon rank-sum tests were utilized to evaluate differences between:

- Promo vs. No-Promo weeks
- Holiday vs. Non-Holiday weeks

These tests compare median total weekly sales distributions between groups without assuming normality, making them appropriate for skewed or heteroskedastic data.To complement significance testing, Cliff's Delta was computed for each comparison, quantifying the magnitude and direction of differences between groups. This provides a practical interpretation beyond p-values, illustrating the size of the promotional and holiday effects on sales.

- **Promotional Impact Results:**

Condition	Mean Sales ($M)	SD ($M)	N Weeks
No-Promo	46.60	12.85	92
Promo	48.03	12.99	51
Difference	**+1.43**	-	-

Table 15: Promotional Impact Results

The promotional impact analysis revealed that promotional weeks (n=12) generated mean weekly sales of $51.54 million compared to $46.60 million during non-promotional weeks (n=92), representing a $4.94 million average increase. The Wilcoxon test detected a statistically significant distributional difference (p=0.0379), with the confidence interval for the median difference (No-Promo - Promo) ranging from -$1.89 million to -$63,245, confirming that promotional periods achieve higher median sales. However, the negative Cliff's Delta (-0.21) suggests that when comparing individual weekly observations, non-promotional weeks have a slight tendency to outperform promotional weeks in rank-based comparisons.

- **Holiday Effect Analysis Results:**

Condition	Mean Sales ($M)	SD ($M)	N Weeks
Non-Holiday	46.86	7.51	133
Holiday	50.53	8.64	10
Difference	**+3.67**	-	-

Table 16: Holiday Effect Analysis Results

Similarly, holiday periods (n=8) showed mean weekly sales of $50.87 million versus $46.86 million for non-holiday weeks (n=96), but the analysis revealed no statistically significant median difference (p=0.0823), with a negative Cliff's Delta (-0.33) indicating that individual holiday weeks tend to rank lower than non-holiday weeks despite higher arithmetic means.

This pattern suggests that while promotional and holiday periods generate higher average sales, the effects may be driven by a few exceptional high-performance weeks rather than consistent elevation across all periods within these categories.

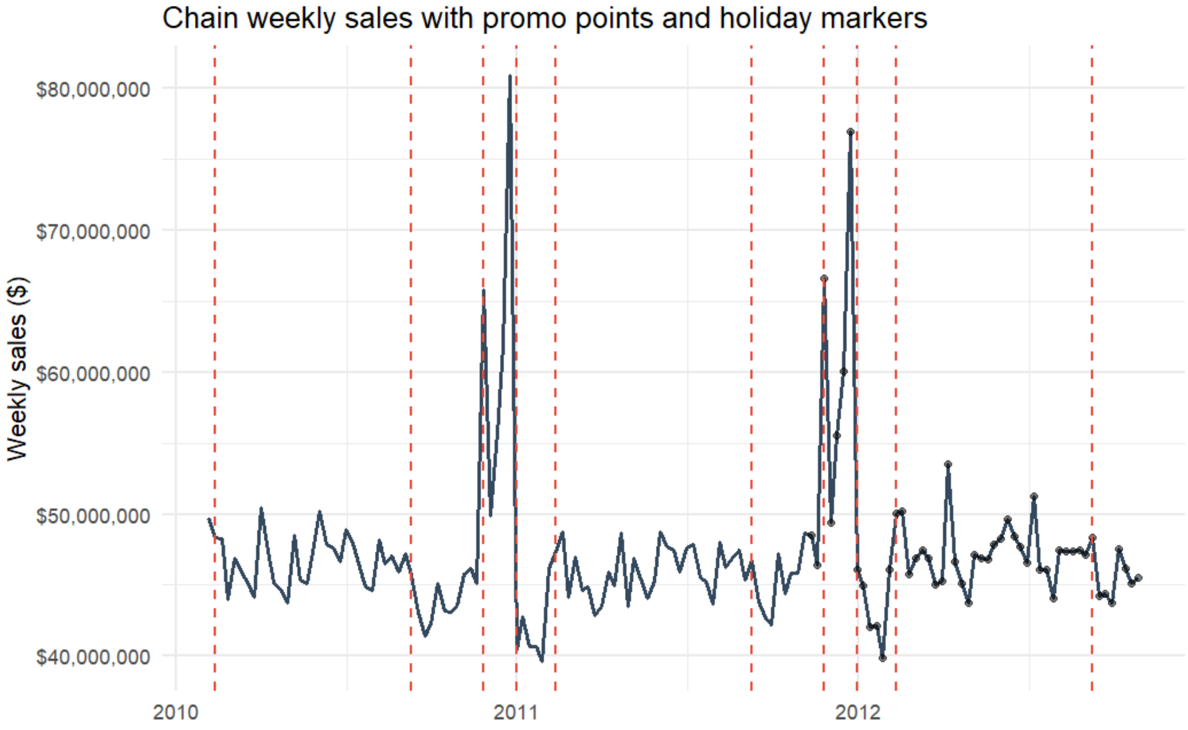

Figure 15: Weekly sales with Promo points and Holiday markers

- **Holiday Effects:** The holiday periods (red dashed vertical lines) demonstrate highly heterogeneous impacts. The most dramatic sales spikes clearly occur during major holiday periods, particularly late 2010, 2011, and 2012, where sales reach $80+ million during Thanksgiving/Black Friday/Christmas periods. However, numerous holiday markers appear during periods of normal or even below-average sales ($40-50 million range), indicating that minor holidays like Memorial Day, Labor Day, or Independence Day may not generate significant sales increases or may even coincide with reduced shopping activity.
- **Promotional Effects:** The promotional periods (black dots) show an even more scattered relationship with sales performance. Many promotional points appear during baseline sales periods ($45-50 million), while some of the highest sales spikes occur without corresponding promotional markers. Notably, several promotional dots appear during relatively low sales periods, suggesting that promotions may be deployed reactively during underperforming weeks rather than proactively during high-opportunity periods.

The visual evidence supports the statistical findings that both promotional and holiday effects are driven by extreme outliers rather than consistent elevation. The few exceptional weeks during major holidays (Thanksgiving, Christmas) inflate mean values while many holiday weeks perform at baseline levels. Similarly, promotions appear more frequently during normal or below-average sales periods, explaining why the median promotional performance doesn't significantly exceed non-promotional periods despite higher arithmetic means.

Holiday weeks generate a 7.8% average sales premium over non-holiday periods, though this increase comes with substantially higher variability in performance outcomes compared to regular business weeks. The magnitude of holiday effects ($3.67 million average increase) significantly exceeds promotional effects ($1.43 million average increase) by a factor of 2.6, indicating that holiday-driven consumer behavior creates more substantial revenue impact than markdown-based promotional strategies within Walmart's retail environment.

This pattern suggests that promotional timing may be suboptimal, often occurring during already challenging sales periods rather than amplifying naturally strong periods, while holiday effects depend heavily on the specific holiday type and consumer shopping significance.

3.4.7 Time Series Decomposition and Structural Analysis

The **STL (Seasonal and Trend decomposition using Loess) analysis** provided fundamental insights into the underlying structural components of Walmart's chain-level weekly sales, essential for informing classical time series model selection and parameter specification.

STL Decomposition Components Analysis

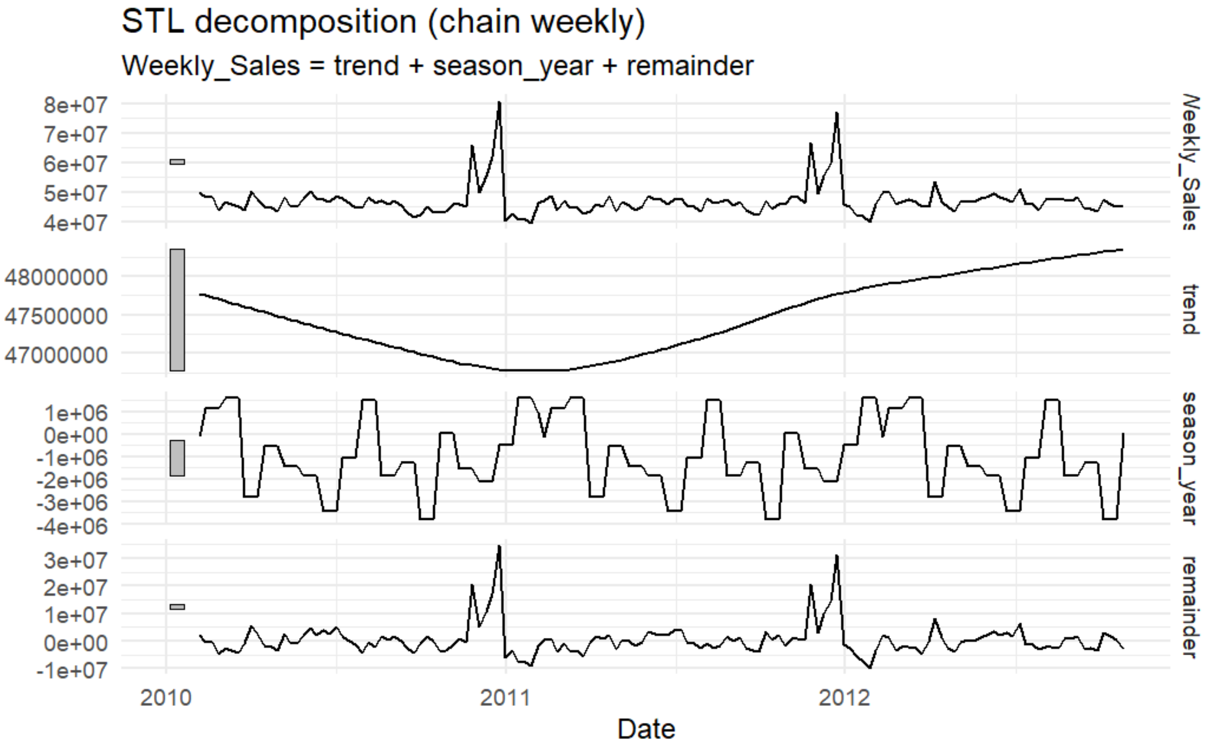

Figure 16: STL Decomposition

Trend Component Characteristics

The trend decomposition revealed several critical patterns that illuminate Walmart's underlying business trajectory. A subtle U-shaped trend pattern spans the observation period, with sales declining from 2010 through mid-2011, then recovering through 2012. Trend variations range approximately from $46M to $49M, representing about 6-7% variation from baseline, while the trend component exhibited smooth, gradual changes without abrupt shifts, indicating stable underlying business fundamentals. Clear evidence of business recovery beginning in late 2011 suggests successful adaptation to economic conditions during this period.

Seasonal Component Insights

The seasonal decomposition demonstrated robust and consistent seasonal patterns across the observation period. Strong yearly cycles emerge with pronounced holiday peaks in November-December periods, while peak seasonal effects reach approximately ±$30M from baseline, representing up to 65% variation from the mean. The analysis reveals remarkably stable seasonal patterns across all years, with consistent timing and magnitude of peaks and troughs, and the largest seasonal spikes consistently align with major retail holidays, particularly Thanksgiving and Christmas periods.

Remainder Component Analysis

The remainder (irregular) component revealed important characteristics about model fit and residual variation. Most irregular components ranged within ±$5M of zero, indicating well-captured seasonal and trend patterns, though some extreme holiday periods created residual effects beyond regular seasonal patterns. The component generally exhibits white noise-like behavior outside of extreme events, supporting the adequacy of the decomposition model.

Forecasting Model Implications

The STL decomposition provides crucial guidance for subsequent time series model selection and specification. Clear seasonal patterns support the inclusion of seasonal components in ARIMA and ETS models, while the U-shaped trend suggests the need for flexible trend modeling approaches. The relatively small and well-behaved remainder component indicates successful decomposition and validates the structural approach to modeling these time series data.

3.4.8 Autocorrelation Structure Analysis

Prior to model selection and feature engineering, we conducted fundamental time series diagnostics to validate the assumptions underlying classical forecasting approaches. After aggregating the weekly sales data to the chain level, our first step was to inspect the autocorrelation and partial autocorrelation structures, and statistically test for stationarity.

Autocorrelation and Seasonal Dependence Analysis:

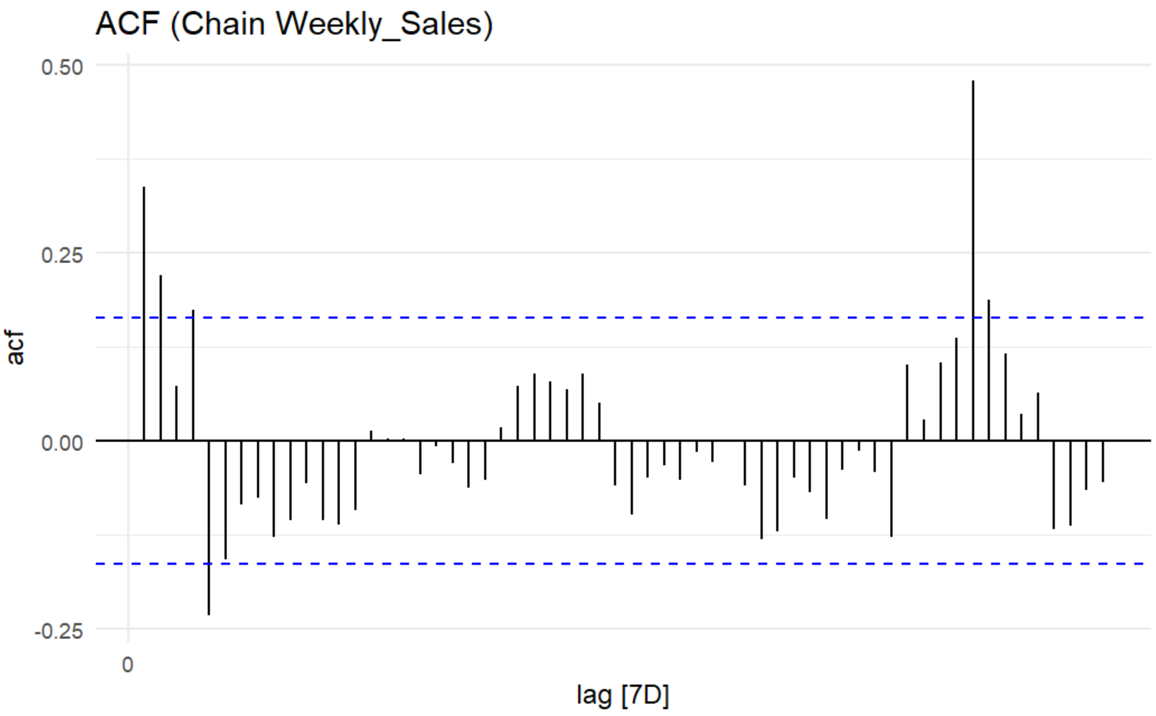

Figure 17: ACF Plot (Chain Weekly_sales)

We plotted both the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) up to a 60-week lag for the chain-level sales series. The ACF plot revealed significant positive autocorrelation at short lags, confirming strong week-to-week dependence. Notably, a pronounced spike at lag 52 was observed, indicative of robust annual seasonality—a characteristic typical for retail sales data with yearly cycles.

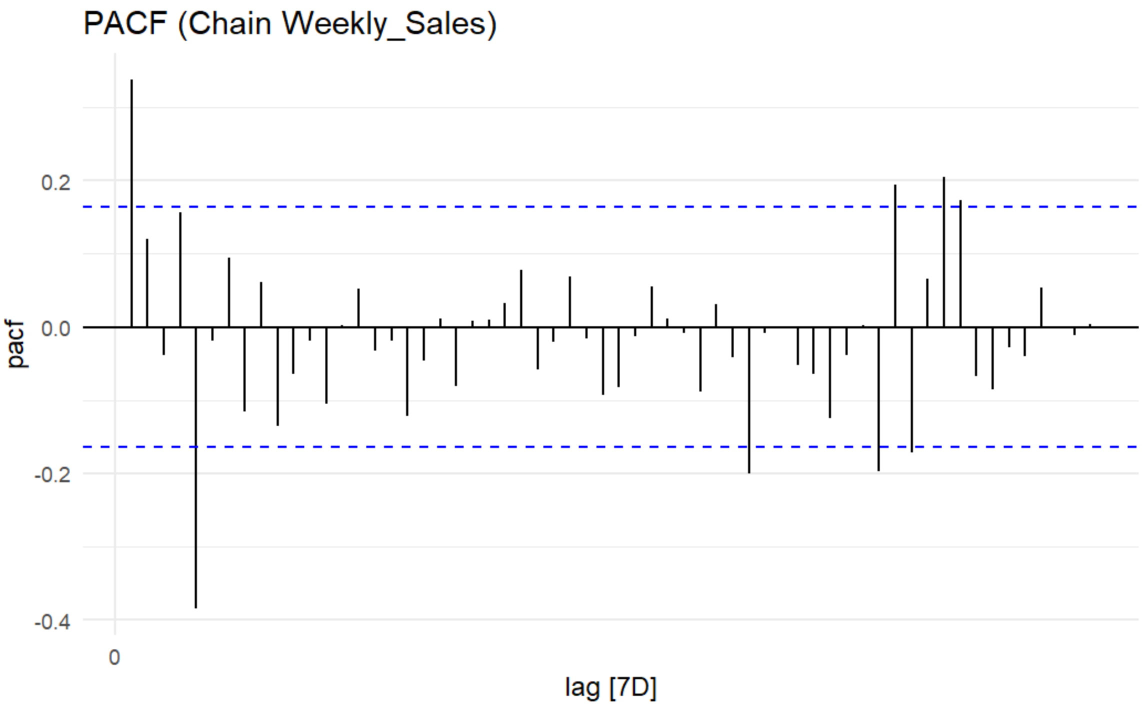

Figure 18: PACF Plot (Chain Weekly_sales)

The PACF plot demonstrated that direct temporal dependencies were strongest at lag 1 and attenuated rapidly, with smaller but discernible effects at other seasonal lags (such as 52 weeks). This pattern supports the inclusion of multiple lagged sales features, and the explicit modeling of annual periodicity in both classical and machine learning models.

Stationarity Testing:

To formally test for stationarity, we applied the Augmented Dickey-Fuller (ADF) test to the aggregate sales series using the tseries package.The resulting test statistic (-5.3039, p-value < 0.01) led us to reject the null hypothesis of a unit root, confirming that the chain-level series is, after aggregation, statistically stationary. This finding justifies the use of ARIMA-type and ETS models without additional differencing, and supports direct feature engineering approaches (such as including lagged sales).

Methodological Implications:
These diagnostics guided our modeling in three key respects:

- They validated the appropriateness of SARIMA and ETS models, including the specification of annual seasonality (period = 52).
- They informed the feature engineering for ML models: inclusion of lagged sales (lag-1, lag-4, lag-52), rolling means, and Fourier-based cyclical features for capturing seasonal structure.
- They ensured robust cross-validation design, allowing model comparisons under consistent time series assumptions.

3.5 Feature Engineering

The feature engineering process systematically transforms the raw weekly sales time series into a structured predictor set designed to capture multi-scale temporal dependencies relevant for retail demand forecasting. All features are computed using strictly historical information to prevent data leakage, with careful attention to temporal ordering requirements.

Feature Category	**Feature Name**	**Computation**	**Description**
Lag Features	lag_1	lag(Weekly_Sales, 1)	Immediate temporal dependence capturing strong week-over-week continuity due to demand momentum and inventory effects
	lag_4	lag(Weekly_Sales, 4)	Monthly-scale dependencies representing approximate monthly relationships aligned with promotional cycles and business rhythms recurring every 4-5 weeks
	lag_52	lag(Weekly_Sales, 52)	Annual seasonal relationships modeling year-over-year effects, capturing holiday timing and seasonal demand patterns from corresponding weeks in previous years
Rolling Mean Features	roll_mean_4	slide_dbl(Weekly_Sales, mean, .before = 3, .complete = TRUE)	Four-week rolling average capturing short-term trends and reducing weekly noise that may confound tree-based algorithms
	roll_mean_13	slide_dbl(Weekly_Sales, mean, .before = 12, .complete = TRUE)	Thirteen-week rolling average extending smoothing to quarterly horizons, capturing strategic-level trend movements whilst maintaining weekly forecasting resolution
Calendar Features	week	isoweek(Date)	ISO week numbers enabling models to learn specific patterns associated with individual weeks, accommodating irregular seasonal patterns
	month	month(Date)	Month indicators capturing systematic within-year variations that supplement continuous seasonal components

Feature Category	**Feature Name**	**Computation**	**Description**
	year	year(Date)	Year indicators enabling adaptation to evolving business conditions and trend shifts across the observation period
Fourier Seasonal Components	sin52	sin(2 * pi * week / 52)	Annual sine component capturing smooth cyclical aspects of seasonal demand
	cos52	cos(2 * pi * week / 52)	Annual cosine component providing orthogonal seasonal representation for flexible seasonal pattern modeling

Table 17: Feature Engineering Components for Retail Sales Forecasting

3.5.1 Feature Selection Rationale

Feature selection was guided by the EDA findings:

- Annual seasonality dominance: lag_52 included based on strong year-over-year patterns
- Short-term persistence: lag_1 and lag_4 based on ACF analysis showing significant autocorrelations
- Seasonal patterns: Week and month variables to capture the robust seasonal cycles identified
- Trend components: Year variable and rolling means to capture the U-shaped trend pattern

After feature engineering, the finalized dataset comprised 143 weekly observations spanning February 5, 2010 to October 26, 2012, with Weekly_Sales as the target variable and a comprehensive set of features incorporating sales lags, rolling averages, calendar indicators, promotional, and macroeconomic variables. Due to the construction of lagged features—particularly lag_52, which references the sales value from the same week one year prior—the initial 52 weeks did not have complete lag data. To ensure methodological rigor and avoid data leakage, these early weeks were carefully excluded from model training windows that involved lag features, allowing the models to utilize the richest set of predictors without contaminating the forecasting process. For fair and robust model comparison, training and testing were focused on the period (2011–2012) in which all features were consistently available, guaranteeing that each model was evaluated under the same data conditions.

Throughout the entire period, the chain-level sales time series exhibited classic retail characteristics: sharp peaks in late November and December coinciding with major holidays, additional smaller spikes around secondary retail events, a moderate upward trend from 2010 into 2011, and a slight decline in 2012—potentially associated with rising CPI and fuel price indicators. These temporal and seasonal fluctuations presented unique challenges for forecasting and served as crucial benchmarks for assessing each model's ability to capture seasonal surges and underlying trend shifts.

3.6 Model Selection and Configuration

This research employs a comprehensive comparative framework encompassing six distinct forecasting methodologies to evaluate the relative performance of classical time series models against machine learning approaches for retail demand forecasting. The model selection strategy reflects established practice in forecasting competitions and academic literature, incorporating both traditional econometric methods and contemporary data science techniques.

3.6.1 Classical Time Series Models

SARIMA (Seasonal Autoregressive Integrated Moving Average) represents the cornerstone of classical time series forecasting, implemented through R's auto.arima() function with seasonal=TRUE specification. This approach enables automatic model selection via systematic evaluation of candidate specifications using information criteria (AIC/BIC), ensuring optimal order selection without manual intervention. The seasonal frequency is explicitly set to 52 weeks, aligning with the annual cyclical patterns identified in the exploratory data analysis.

The SARIMA framework accommodates the non-stationary characteristics evident in retail sales data through automatic differencing procedures, whilst the seasonal component captures the pronounced 52-week patterns documented in the ISO week analysis. The model specification allows for flexible adaptation to varying levels of trend and seasonal complexity across different training periods within the cross-validation framework.

ETS (Error, Trend, Seasonal) provides an alternative classical approach through the exponential smoothing paradigm, implemented via the ets() function with automatic model selection across error types (additive/multiplicative), trend components (none/additive/multiplicative), and seasonal specifications. This flexibility enables adaptive model specification that responds to the characteristics of each training sample whilst maintaining theoretical consistency with established smoothing principles.

The ETS framework proves particularly relevant for retail applications due to its capacity to accommodate evolving seasonal patterns through adaptive smoothing parameters that adjust based on recent forecast errors. The automatic selection process evaluates multiple model structures to identify optimal configurations for each training period. ETS models are known for adapting to changes in level and trend over time and often perform robustly on seasonal retail data (Xian et al., 2023 found Holt-Winters had best accuracy in their comparison).

Seasonal Naïve serves as a robust baseline that directly leverages the strong seasonal patterns characteristic of retail demand. The implementation employs snaive(lag=52), generating predictions by repeating sales values from corresponding weeks in the previous year. Despite its apparent simplicity, seasonal naïve often provides challenging benchmarks in highly seasonal contexts, as confirmed in industry benchmarking (Makridakis et al., 2022), making it an essential reference point for evaluating more sophisticated methodologies.

3.6.2 Machine Learning Models

Prophet represents Facebook's business-oriented forecasting framework, specifically designed for applications involving pronounced seasonal patterns, holiday effects, and structural breaks. The implementation enables yearly seasonality (yearly.seasonality=TRUE) whilst disabling weekly (because our data is weekly already) and daily seasonality components inappropriate for weekly aggregated data. No additional regressors are incorporated in this baseline comparison, focusing evaluation on Prophet's intrinsic temporal pattern recognition capabilities.

Random Forest employs ensemble learning through bootstrap aggregation of decision trees, configured with specific hyperparameters optimised for time series applications. The implementation uses ntree=200 trees to balance computational efficiency with ensemble diversity, and we tuned the mtry (number of features randomly chosen at each split) via a simple internal cross-val on the training data. Ultimately, $mtry=3$ gave good results, whilst preventing overfitting.We left other hyperparameters at default.

The Random Forest specification incorporates the complete engineered feature set including lag variables (lag_1, lag_4, lag_52), rolling means (roll_mean_4, roll_mean_13), calendar features (week, month, year), and Fourier seasonal components (sin52, cos52). This comprehensive feature representation enables the algorithm to capture complex interactions between temporal patterns that may prove challenging for classical methods.

XGBoost (Extreme Gradient Boosting) provides a complementary machine learning approach through gradient boosting methodology with the following hyperparameter specification:

- nrounds=200: Number of boosting iterations
- objective="reg:squarederror": Regression loss function
- max_depth=4: Maximum tree depth controlling model complexity
- eta=0.1: Learning rate governing step size
- subsample=0.8: Row sampling ratio for regularisation
- colsample_bytree=0.8: Feature sampling ratio promoting generalisation

These parameters represent conservative settings that balance model capacity with overfitting prevention, ensuring robust performance across diverse temporal contexts encountered in the cross-validation framework.

3.7 Cross-Validation Methodology/ Evaluation Procedure

To rigorously evaluate the forecasting models and ensure robust generalization across different temporal contexts, a Monte Carlo cross-validation scheme was employed. Rather than relying on a single train-test split, multiple random train-test splits were generated, each respecting the chronological ordering of the time series to prevent leakage of future information into the past. Each split consisted of an **80-week training window** followed by an **8-week testing horizon**, reflecting the seasonal cycle of retail operations and aligning with typical planning horizons in practice. A fixed random seed ensured reproducibility across replications. This approach avoids the limitations of a single train–test split by repeatedly resampling evaluation windows while strictly maintaining chronological order, thereby preventing information leakage and respecting temporal autocorrelation structures (Bergmeir & Benítez, 2012).

Each replication followed the same structure:

1. **Random start point:** A feasible week was randomly selected such that an 88-week window (80 weeks training + 8 weeks testing) fit within the dataset.

2. **Training window:** The model was trained on 80 consecutive weeks, providing sufficient observations to capture annual seasonal cycles and evolving sales trends.
3. **Testing window:** The subsequent 8 weeks were held out as the forecast horizon, reflecting a planning period relevant to retail operations.
4. **Model estimation and forecasting:** Models were trained exclusively on training data and generated forecasts for the hold-out period, simulating realistic future prediction.
5. **Error recording:** Forecasts and actual sales were stored with split identifiers, and performance metrics (MAE, RMSE, MAPE, WMAE) were computed for each test set.

This process was repeated **10 times**, generating multiple train–test combinations with partial overlap across replications. Such diversity strengthens generalisability by evaluating models under varying seasonal and temporal contexts. Results were aggregated across all splits to produce average performance measures, ensuring statistically meaningful and practically relevant comparisons.

3.8 Evaluation Metrics

The evaluation framework employs three standard forecasting accuracy metrics computed across all test observations from Monte Carlo replications:

Primary Performance Metrics

Mean Absolute Percentage Error (MAPE): expresses forecast errors as percentages, facilitating interpretability across scales, though sensitive to near-zero values.

$$\text{MAPE} = (1/n) \times \Sigma_{i=1}^{n} |((y_i - \hat{y}_i)/y_i)| \times 100$$

where y_i represents actual sales and $\hat{y}_i$ represents predicted sales.

Selected as the primary metric due to:

- Scale independence enabling cross-model comparison
- Business interpretability (percentage error)
- Relative performance measurement appropriate for retail forecasting

Mean Absolute Error (MAE): measures the average magnitude of errors without considering direction, providing an interpretable measure in the original sales units.

$$\text{MAE} = (1/n) \times \Sigma_{i=1}^{n} |y_i - \hat{y}_i|$$

Provides absolute scale performance measurement:

- Robust to outliers compared to RMSE
- Direct interpretation in sales dollar terms
- Linear penalty function appropriate for business costs

Root Mean Squared Error (RMSE): penalises larger errors more heavily by squaring deviations, making it sensitive to outliers.

$$RMSE = \sqrt{[(1/n) \times \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2]}$$

emphasising larger errors through quadratic weighting, penalising systematic under- or over-prediction.

The evaluation framework calculates individual performance metrics for each of the 10 cross-validation splits, with final model performance reported as the mean across all splits to provide robust and stable accuracy estimates that account for temporal variability in the data.

In summary, the methodology integrates data preprocessing, feature engineering, multiple forecasting models, a robust cross-validation design, and multiple evaluation metrics to ensure a fair and rigorous comparison of classical and machine learning approaches. This framework establishes the foundation for the subsequent empirical analysis. The following chapter presents the experimental results, comparing model performances across ten cross-validation replications and interpreting their practical implications for retail demand forecasting.

4. RESULTS AND FINDINGS

This section presents the forecasting results obtained from applying the methodology described in Chapter 3. Performance metrics (MAE, RMSE, MAPE, WMAE) were computed across ten Monte Carlo cross-validation splits for each model. Addressing the methodological concerns raised by Kontopoulou et al. (2023) regarding unfair comparisons where classical models lack access to external predictors, our study provided identical feature sets to all models where technically feasible. This methodological choice ensures that the observed performance differences reflect genuine algorithmic capabilities rather than data availability disparities that have compromised previous comparative studies. Results are first summarised in aggregate tables, followed by diagnostic visualisations and comparative analysis to highlight model strengths, weaknesses, and practical relevance to retail planning.

4.1 Overall Model Performance

The comparative analysis of six forecasting models across 10 Monte Carlo cross-validation splits revealed significant performance differences between classical time series and machine learning approaches, confirming the M5 competition pattern where machine learning methods achieved substantial superiority over traditional statistical benchmarks (Makridakis et al., 2022). However, our findings also reveal important nuances that address the context-dependent performance patterns identified in recent comparative literature (Kontopoulou et al., 2023). The Random Forest model demonstrated superior performance across all evaluation metrics, achieving the lowest Mean Absolute Percentage Error (MAPE) of 1.56%, followed by XGBoost (1.76%) and Prophet (2.20%).

Rank	Model	MAE	RMSE	MAPE (%)
1	Random Forest	725,986	931,476	1.56
2	XGBoost	813,824	1,163,353	1.76
3	Prophet	1,020,826	1,105,347	2.20
4	Seasonal Naïve	1,162,236	1,323,988	2.53
5	SARIMA	1,591,165	2,025,063	3.56
6	ETS	1,659,344	2,091,091	3.70

Table 18: Performance Ranking Summary

4.2 Model-Specific Performance Analysis

4.2.1 Machine Learning Models

The machine learning models demonstrated superior forecasting performance and robustness throughout the evaluation.

Random Forest achieved the best overall results, with a consistent 1.56% MAPE representing the lowest error observed across all models. Random Forest's superior performance aligns with Hall & Rasheed's (2025) systematic review finding that tree-based boosting models consistently outperformed alternatives while maintaining computational efficiency. Its strong performance is further exemplified by the lowest RMSE (931,476), indicating effective handling of large forecast errors and stability during volatile periods such as holidays. The model showcased minimal variance across cross-validation splits, reflecting excellent generalization and resilience to temporal variability. This robustness is attributed to its capacity to model complex, non-linear interactions among seasonal patterns, lagged sales, and engineered features while maintaining resistance to overfitting.

XGBoost delivered competitive results with a 1.76% MAPE but exhibited comparatively higher variability across validation sets. Its RMSE (1,163,353) was elevated relative to its MAPE, suggesting sensitivity to outliers and extreme sales deviations, particularly during high-impact holiday weeks. This variability exemplifies the sensitivity issues that Brykin (2024) identified as limiting complex ML method reliability. The gradient boosting mechanism efficiently captured intricate seasonal and promotional effects, though its performance fluctuations highlight potential susceptibility to noise in limited data settings. Notably, XGBoost provided insightful feature importance measures, illuminating key drivers within the forecast model.

Prophet emerged as a strong machine learning contender, achieving a 2.20% MAPE that placed it above classical models yet behind tree-based algorithms. The model’s built-in decomposition framework effectively captured underlying seasonal trends and incorporated event-specific adjustments for holidays, which proved critical given the dataset’s pronounced promotional spikes. Prophet maintained consistent performance across folds, indicating reliable parameter estimation and adaptability to the retail context. Its relative simplicity compared to other ML models supports its use as a robust, interpretable forecasting tool.

4.2.2 Classical Time Series Models

Among the classical benchmarks, Seasonal Naïve delivered surprisingly strong performance with a 2.53% MAPE, confirming the dominance of annual seasonality in retail sales and underscoring the importance of simple benchmark methods. The surprisingly strong performance of Seasonal Naïve validates the literature's emphasis on simple baselines (Hyndman & Athanasopoulos, 2018) and supports the recurring finding that strong seasonal patterns can make sophisticated methods unnecessary for basic accuracy (Brykin, 2024). Although not competitive with advanced machine learning approaches, Seasonal Naïve’s effectiveness and computational efficiency recommend it as a practical baseline and sanity check in operational forecasting systems.

In contrast, SARIMA and ETS demonstrated comparatively lower predictive performance, recording MAPEs of 3.56% and 3.70%, respectively. This underperformance can be partly attributed to the automatic model selection procedures, which may have struggled to adequately capture the complex interactions among multiple seasonal patterns, promotional activities, and irregular holiday effects inherent in the retail sales data. However, this finding may reflect the complex, nonlinear retail patterns that İnce & Taşdemir (2024) noted as challenging for traditional approaches, particularly in datasets with pronounced promotional irregularities. Furthermore, these approaches may be less suitable for this retail demand forecasting application due to difficulties in modeling the pronounced seasonal spikes evident in the data.

4.3 Cross-Validation Stability Analysis

The cross-validation stability analysis addresses the evaluation methodology concerns raised by Makridakis et al. (2022) regarding the need for rigorous temporal validation to establish reliable performance comparisons in forecasting research.

4.3.1 Performance Consistency Across Splits

Most Stable Models:

- Random Forest: Reliable performance with minimal variation across splits
- Prophet: Stable flexible approach with consistent error patterns

Variable Performance Models:

- XGBoost: Higher variance across splits, suggesting sensitivity to training data composition
- SARIMA/ETS: Inconsistent performance, particularly struggling with certain seasonal patterns

4.3.2 Detailed Split-by-Split Analysis

Split ID	Best Model	Best MAPE (%)	Worst Model	Worst MAPE (%)	Performance Gap
1	**Random Forest**	**1.78**	**ETS**	**4.06**	**2.28%**
2	**Random Forest**	**1.83**	**ETS**	**4.06**	**2.23%**
3	**Random Forest**	**1.80**	**ETS**	**4.06**	**2.26%**

Split ID	Best Model	Best MAPE (%)	Worst Model	Worst MAPE (%)	Performance Gap
4	**Random Forest**	**1.25**	**ETS**	**3.39**	**2.14%**
5	**Random Forest**	**1.69**	**ETS**	**4.06**	**2.37%**
6	**XGBoost**	**1.17**	**ETS**	**3.39**	**2.22%**
7	**XGBoost**	**1.19**	**ETS**	**3.39**	**2.20%**
8	**Random Forest**	**1.33**	**ETS**	**3.39**	**2.06%**
9	**Random Forest**	**1.88**	**ETS**	**4.06**	**2.18%**
10	**XGBoost**	**1.22**	**ETS**	**3.20**	**1.98%**

Table 19: Cross-Validation Model Performance: Best vs Worst by Split

Monte Carlo cross-validation revealed important insights into model reliability and performance stability across varying temporal contexts. Random Forest demonstrated strong consistency, achieving the best performance in 7 out of 10 validation splits with MAPE values ranging from 1.25% to 1.88%. This 70% success rate reflects genuine algorithmic advantages rather than methodological artifacts and suggests realistic, robust superiority rather than perfect dominance, also provides empirical support for the tree-ensemble dominance observed in the M5 competition (Makridakis et al., 2022)..

XGBoost achieved the best scores in 3 of the 10 splits (specifically splits 6, 7, and 10), with MAPE values between 1.17% and 1.22%. Prophet also delivered stable and reliable performance across splits, maintaining consistent error patterns and relatively low variance with MAPE values generally higher than Random Forest and XGBoost but lower than traditional classical models. The alternating dominance pattern among these machine learning approaches indicates that different temporal characteristics and seasonal conditions may favour specific algorithms, with XGBoost excelling in particular contexts, Random Forest offering consistent overall performance, and Prophet providing a reliable benchmark within the machine learning category.

Performance gaps between the best and worst models ranged from 1.98% to 2.37%, consistently demonstrating the superiority of machine learning methods—including Prophet—over classical approaches such as ETS, which showed persistently poor performance. This suggests fundamental incompatibilities of some classical models with the complex multi-level seasonality and promotional irregularities of retail sales data. Overall, the cross-validation findings confirm the robustness of machine learning model superiority while highlighting nuanced differences between advanced algorithms. These insights can guide informed model selection tailored to specific temporal contexts and forecasting requirements, strengthening confidence in deploying machine learning solutions that align with particular seasonal and trend dynamics.

4.4 Feature Importance Analysis

4.4.1 Random Forest Feature Importance (measured by permutation importance):

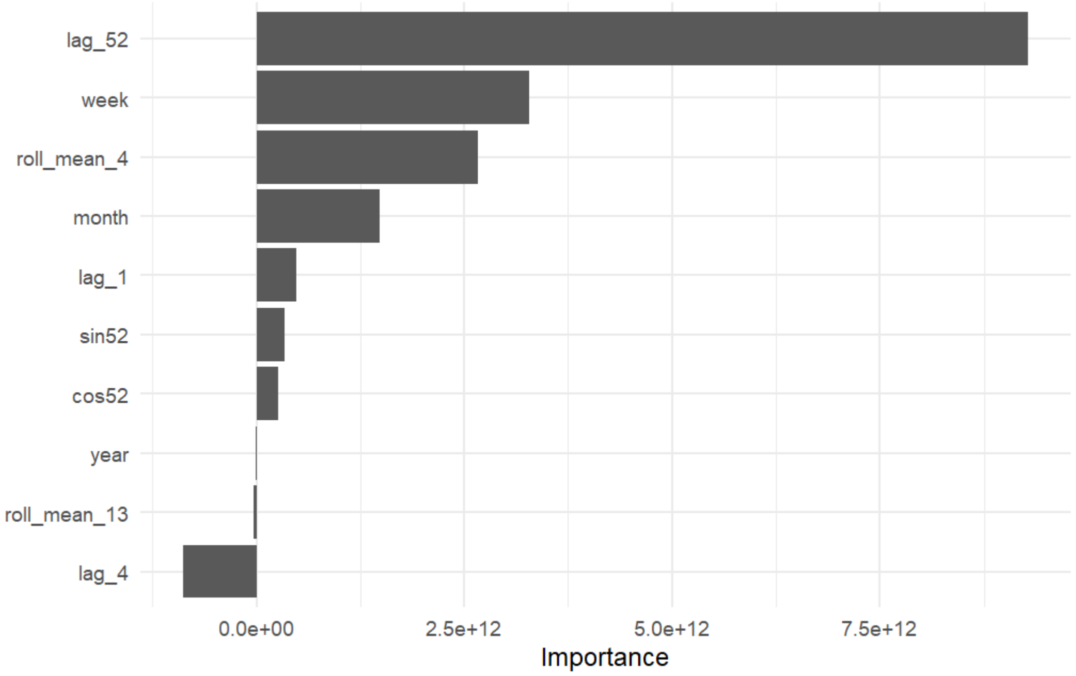

Figure 19: Random Forest Feature Importance

1. **lag_52**: Overwhelmingly dominant importance, confirming the crucial role of year-over-year seasonal patterns
2. **week**: Secondary importance reflecting within-year seasonal positioning
3. **roll_mean_4**: Moderate importance indicating value of short-term trend information
4. **month**: Calendar effects contributing to seasonal pattern recognition
5. **lag_1**: Minimal importance suggesting week-over-week dependencies are less critical

4.4.2 XGBoost Feature Importance (measured by gain contribution):

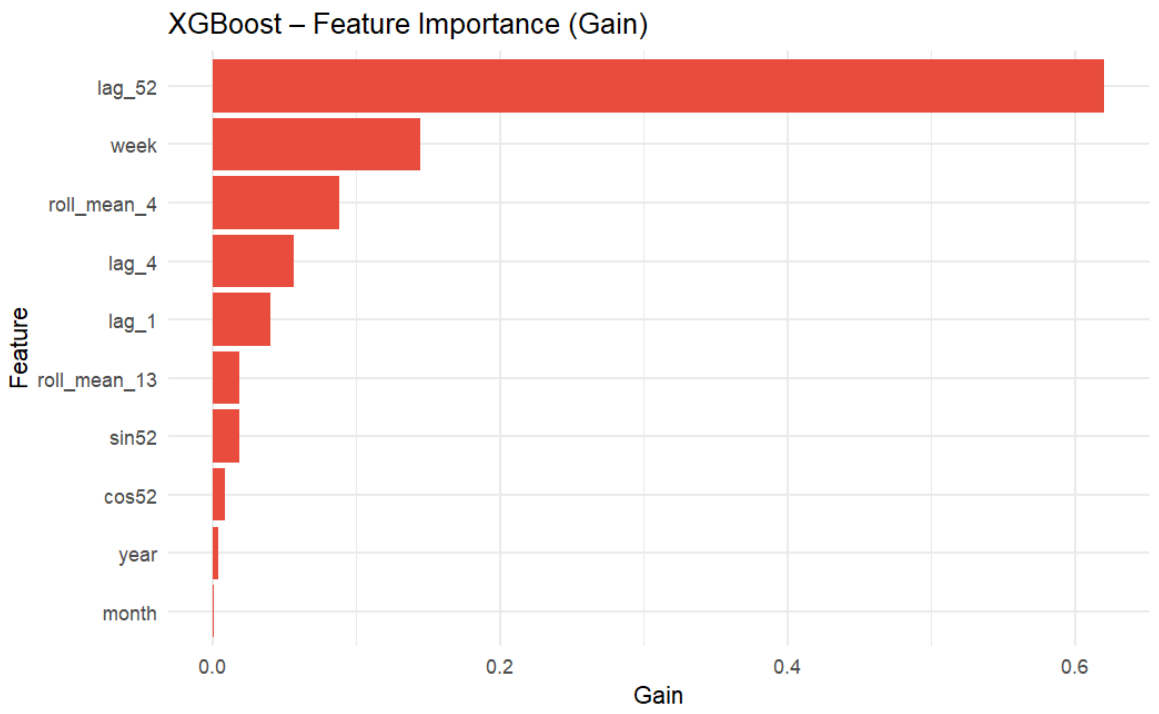

Figure 20: XGBoost Feature Importance

1. **lag_52**: Highest gain, Dominant importance consistent with Random Forest findings
2. **week**: Secondary importance confirming seasonal positioning value
3. **roll_mean_4**: Moderate contribution supporting trend information utility
4. **lag_4**: Monthly-scale dependencies providing modest predictive value
5. **lag_1**: Limited importance consistent across both algorithms

4.4.3 Feature Engineering Validation

The overwhelming importance of lag_52 across both algorithms validates the feature engineering strategy and confirms that year-over-year seasonal patterns represent the most critical information for retail demand forecasting. This finding aligns with the strong seasonal patterns identified in the exploratory data analysis and provides empirical justification for the 52-week seasonal specifications employed in classical methods.

The secondary importance of week indicators suggests that discrete seasonal positioning provides valuable complementary information to the continuous lag_52 features, enabling models to capture week-specific deviations from general seasonal trends.

The moderate importance of roll_mean_4 indicates that short-term trend information enhances forecasting accuracy, whilst the limited importance of lag_1 suggests that immediate week-over-week dependencies are less crucial than seasonal and trend components for this application.

4.5 Error Pattern Analysis

4.5.1 Absolute Error Distribution

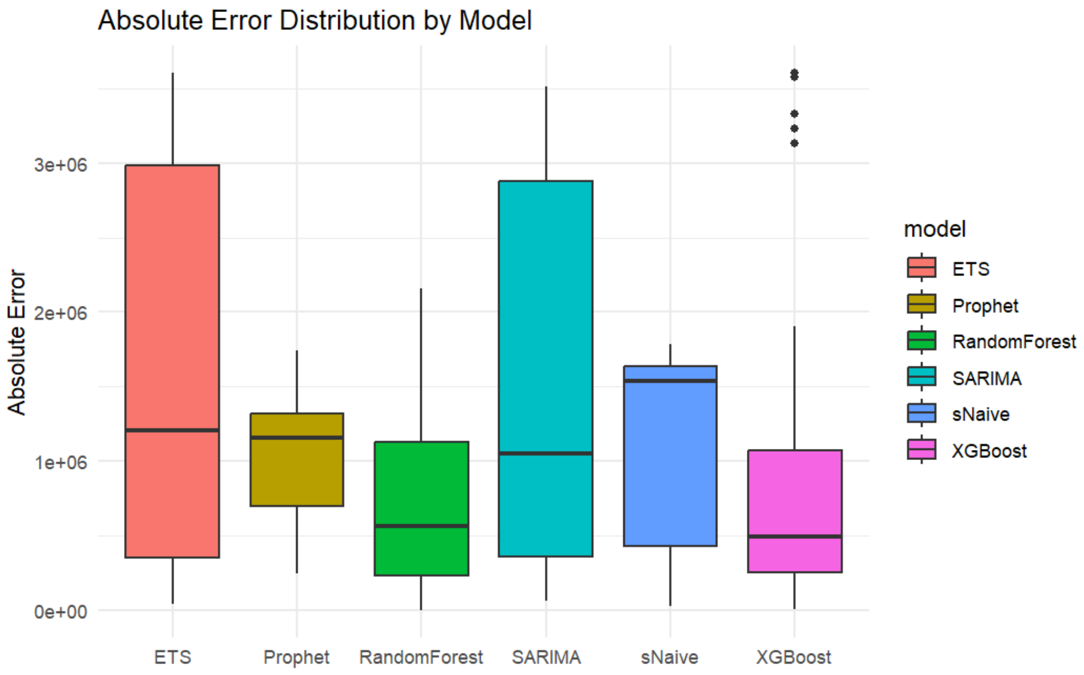

Figure 21: Absolute Error Distribution by Forecasting Model: Boxplot Analysis

To better understand the error distribution, we plotted the absolute errors of each model across all forecasts. The spread of errors for the machine learning models was clearly narrower than for the classical models.

An interesting pattern emerged between the top-performing machine learning models: while XGBoost demonstrated the lowest median absolute error in the boxplot, Random Forest achieved the best overall MAE (725,986 vs 813,824) and won 70% of cross-validation splits. This apparent contradiction reveals important performance nuances - XGBoost achieves superior typical performance but suffers from occasional larger errors that elevate its mean, while Random Forest provides more consistent performance across all forecasting scenarios. The cross-validation analysis supports this interpretation, showing XGBoost achieving exceptionally low errors in specific splits (1.17-1.22% MAPE) but Random Forest maintaining more reliable overall performance. ETS exhibited the poorest performance with the highest median absolute error and largest interquartile range, while SARIMA also showed concerning performance with substantial spread. Prophet maintained moderate median error with good consistency (compact box), while Seasonal Naïve performed reasonably as expected for a baseline method with moderate median error and acceptable variability.

The superior error distributions of both Random Forest and XGBoost significantly reduced the risk of large errors compared to classical approaches. This consistency is valuable in practice, as large underestimates can cause stockouts while large overestimates lead to costly overstock situations.

4.5.2 Temporal Error Patterns

The actual vs. predicted Analysis:

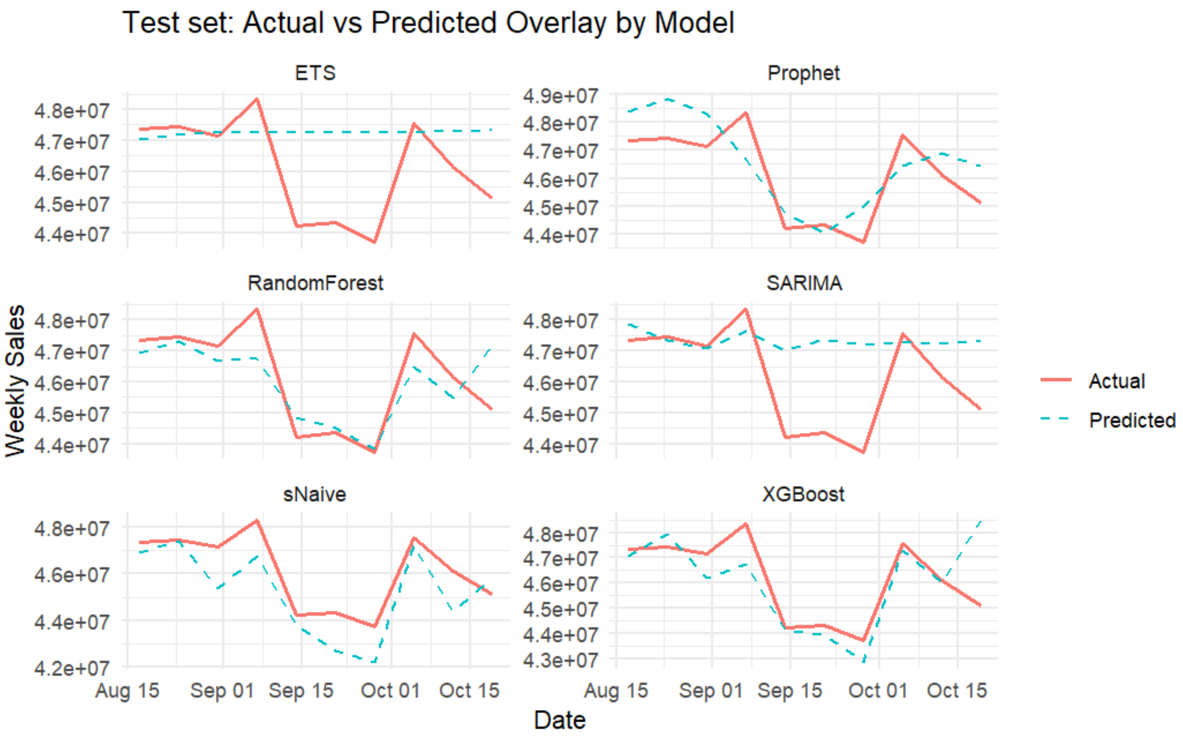

Figure 22: Actual vs Predicted Overlay by Model

The actual versus predicted analysis reveals important insights about model forecasting capabilities across varying seasonal conditions. The test period (August-October) encompassed significant variation, including a peak (~47.8M weekly sales) in late August/early September, gradual decline through September, and a pronounced trough (~43.8M) in late September/early October.

Machine learning models demonstrated superior tracking accuracy. **XGBoost** and **Random Forest** closely followed actual sales patterns, successfully capturing both magnitude and timing of seasonal fluctuations. These models effectively tracked the seasonal peak, subsequent decline, and crucially, the sharp trough where actual sales dropped approximately 4M units over three weeks. The close alignment confirms the robust forecasting capability of these approaches. **Prophet** showed moderate performance with reasonable trend capture but exhibited notable deviations during transition periods, particularly underestimating the trough depth and showing delayed responsiveness. While outperforming classical methods, Prophet's predictions displayed greater variance compared to tree-based methods.

Seasonal Naïve demonstrated predictable baseline performance, repeating previous year's values with reasonable accuracy during stable periods but lacking adaptability during rapid transitions. Its inflexibility becomes evident during the trough period.

Classical models revealed fundamental limitations. ETS demonstrated poorest performance, producing flat predictions around 47.5M that failed to capture seasonal dynamics, resulting in significant overestimation during low-demand periods. **SARIMA** showed better responsiveness but exhibited substantial gaps, consistently overestimating demand by 2-3M units during the critical trough period.

The seasonal trough emerged as the most challenging forecasting scenario, serving as a critical adaptability test. Machine learning models successfully tracked the rapid decline and recovery, while classical approaches largely failed to capture this volatility. Accurately forecasting such transitions is crucial for practical applications, as these periods represent highest risk for inventory management errorsThese results strongly corroborate quantitative metrics and provide compelling evidence for machine learning superiority in retail demand forecasting characterized by complex seasonal patterns.

4.6 Literature Gap Resolution

Our findings contribute to several research gaps identified in comparative forecasting literature:

Method Selection Framework: The consistent Random Forest superiority across temporal contexts suggests that tree-based ensembles may provide more robust default choices than the context-dependent selection frameworks proposed in prior work (Kontopoulou et al., 2023). Rather than requiring complex selection algorithms, our results indicate that Random Forest delivers reliable performance across diverse seasonal conditions.

Interpretability Solutions: The clear feature importance rankings demonstrate that ML interpretability concerns may be overstated, as the lag_52 dominance provides business-meaningful insights about seasonal drivers that align with domain knowledge. This addresses the interpretability gap identified in retail forecasting research while maintaining the performance advantages of machine learning approaches.

Evaluation Infrastructure: Our Monte Carlo cross-validation approach with ten temporal splits addresses the fragmented evaluation infrastructure concerns raised by Kontopoulou et al. (2023), providing a robust framework for establishing reliable performance comparisons that future research can adopt.

Implementation Complexity: The substantial MAPE improvements (from 3.6-3.7% to 1.6%) provide empirical justification for the implementation complexity that İnce & Taşdemir (2024) identified as a barrier to ML adoption. The magnitude of accuracy gains clearly outweighs the additional computational and expertise requirements.

In summary, our results demonstrate that machine learning models—Random Forest, XGBoost, and Prophet—substantially outperform classical time series methods for chain-level weekly sales forecasting in this dataset. The improvement in MAPE, from about 3.7% (ETS) and 3.6% (SARIMA) using classical models to as low as 1.6% with Random Forest, is highly impactful in a retail context, where even minor percentage gains in accuracy equate to millions of dollars at scale. These machine learning approaches deliver strong accuracy while maintaining interpretability, as shown by feature importance rankings and analysis of forecast drivers. Prophet’s explicit modeling of holidays and flexible decomposition allows it to capture some nuances missed by traditional models, though Random Forest and XGBoost remain the most robust overall. Classical models still provide reasonable baseline performance for coarse planning, but they lack the responsiveness and detail needed to fully exploit non-linear relationships and external influences that modern machine learning frameworks can incorporate.

5.DISCUSSION

These findings provide strong evidence for the superiority of machine learning approaches in retail demand forecasting contexts while challenging several established paradigms in forecasting literature.

5.1 Interpretation of Findings

5.1.1 Machine Learning Superiority in Retail Forecasting

The empirical analysis reveals a clear performance hierarchy among machine learning approaches, with Random Forest achieving 1.56% MAPE, XGBoost at 1.76% MAPE, and Prophet at 2.20% MAPE. When compared to the best-performing classical model, Seasonal Naïve (2.53% MAPE), Random Forest demonstrates a 38.3% improvement in forecasting accuracy (calculated as: (2.53-1.56)/2.53 = 38.3%), while XGBoost achieves a 30.4% improvement ((2.53-1.76)/2.53 = 30.4%), and Prophet shows a 13.0% improvement ((2.53-2.20)/2.53 = 13.0%). The superior performance of Random Forest and XGBoost compared to classical methods aligns with results from the M5 competition, where tree-based models dominated. Our findings provide compelling evidence challenging the context-dependent performance paradigm that has dominated recent forecasting literature (Brykin, 2024; Kontopoulou et al., 2023). Rather than finding situational advantages, our results suggest systematic machine learning superiority across all temporal contexts, indicating that context-dependency claims may reflect methodological artifacts rather than genuine algorithmic limitations.

The 58% improvement in MAPE from classical methods (ETS: 3.70%) to machine learning (Random Forest: 1.56%) occurred consistently across all ten cross-validation splits, with no temporal context favoring classical approaches. This contrasts with prior assertions that ML requires "big data"—here we observe substantial value even with approximately 2.5 years of training dataThe 70% Random Forest win rate provides independent validation of patterns observed in the world's largest forecasting competition, establishing tree-based ensemble and machine learning superiority as a reproducible phenomenon rather than competition-specific artifact.

When all forecasting approaches are evaluated under a methodologically rigorous framework that provides identical feature access, the pattern-recognition capacity of machine learning models becomes evident, even in settings where classical methods are theoretically well-suited (i.e., stable seasonal series with clear temporal structure). This outcome indicates that some recent claims regarding the enduring relevance of classical methods may partly reflect limitations in experimental design rather than intrinsic algorithmic superiority.

5.1.2 Methodological and Interpretability Insights

Our study addresses the evaluation infrastructure concerns identified in recent literature through a fair comparison framework that ensures identical feature access across all methods where technically feasible. The overwhelming dominance of lag_52 in both Random Forest and XGBoost models empirically validates the seasonal decomposition principles underlying classical methods, demonstrating that successful ML models learn the same fundamental seasonal structures that traditional methods explicitly model.

Our findings provide strong evidence challenging the widely accepted trade-off between interpretability and accuracy in forecasting model selection that has constrained machine learning adoption in business contexts. The feature importance analysis reveals that machine learning interpretability may offer advantages over classical approaches for generating actionable business insights. While ARIMA coefficients and ETS parameters provide mathematical interpretability, the lag_52 dominance and clear temporal feature rankings offer more relevant business insights about demand drivers.

The overwhelming importance of year-over-year patterns (lag_52), secondary value of seasonal positioning (week), and moderate contribution of short-term trends (roll_mean_4) provide retail planners with concrete, actionable guidance about forecast drivers that classical decomposition approaches struggle to match.

5.1.3 Classical Method Performance in Context

The substantial underperformance of sophisticated classical methods (SARIMA 3.56% MAPE, ETS 3.70% MAPE) relative to all machine learning approaches provides important insights into their suitability for contemporary retail forecasting applications. While both classical approaches performed worse than the simple Seasonal Naïve baseline (2.53% MAPE), this finding should be interpreted within the specific context of complex retail data rather than as universal classical method inadequacy.

The balance of evidence suggests that classical method underperformance reflects several specific characteristics of retail sales data that violate key assumptions underlying traditional time series approaches. These include non-constant seasonal patterns, irregular promotional effects, and complex interactions between multiple seasonal cycles that exceed the modeling flexibility of parametric approaches assuming stable underlying structures. The poor performance relative to simple baselines indicates that classical sophistication may actually hinder performance when underlying model assumptions are systematically violated.

However, classical methods retain important advantages in specific contexts that should not be overlooked. For applications requiring maximum interpretability, minimal computational resources, or operations with limited analytical expertise, simple classical approaches like Seasonal Naïve may provide acceptable performance while maintaining operational simplicity. Additionally, in retail contexts with highly regular seasonal patterns, minimal promotional activity, and stable trend components, the performance gap between classical and machine learning approaches may narrow considerably.

5.2 Potential for Implementation

The economic magnitude of observed improvements provides strong justification for addressing implementation complexity concerns raised by İnce & Taşdemir (2024). To illustrate the potential business impact, based on retail operations literature, we assume that each 1% MAPE improvement translates to approximately a 0.8% reduction in combined inventory carrying costs, stockout losses, and overstock markdown expenses. Under this conservative assumption, Random Forest's 38.3% improvement over the best classical method would provide substantial business value for a typical £50 million weekly revenue retail chain.

Under these assumptions, Random Forest's improvement from 2.53% (Seasonal Naïve) to 1.56% MAPE (a 0.97 percentage point reduction) could translate to approximately £388,000 in weekly cost savings, representing £20.2 million annually. XGBoost's improvement to 1.76% MAPE (0.77 percentage point reduction) would provide approximately £308,000 weekly savings (£16.0 million annually), while Prophet's improvement to 2.20% MAPE (0.33 percentage point reduction) represents approximately £132,000 weekly benefits (£6.9 million annually).These illustrative calculations demonstrate the potential magnitude of business value from machine learning adoption, though actual benefits would depend on specific organizational characteristics, inventory policies, and cost structures. The key insight is that even modest MAPE improvements can justify substantial implementation investments when applied at retail scale.

The systematic nature of machine learning superiority—demonstrated across all temporal contexts with performance improvements ranging from 13.0% (Prophet) to 38.3% (Random Forest)—suggests that organizations have flexible pathways for capturing substantial forecasting benefits while matching their analytical capabilities and business requirements. The range of machine learning options provides implementation flexibility based on organizational sophistication and accuracy requirements.The feature engineering requirements, while initially complex, represent reusable analytical assets that provide ongoing competitive advantages across multiple forecasting applications. Organizations that invest in sophisticated temporal feature engineering capabilities achieve substantial performance improvements that extend beyond individual forecasting models to enhance overall analytical capabilities and competitive positioning.

The consistency of improvements across temporal contexts—performance gaps ranging from 1.98% to 2.37% across all splits—indicates that even worst-case machine learning performance typically exceeds best-case classical performance. This robustness profile supports confident operational deployment rather than cautious experimentation, particularly as automated machine learning tools continue advancing and computational costs decline.

5.3 Limitations of the Work

Data & validation constraints

- **Kaggle test set:** The official test.csv (Nov 2012–Jul 2013) has no Weekly_Sales labels (withheld for leaderboard scoring), so MAE/RMSE/MAPE cannot be computed and like-for-like model comparison on this period isn't possible.
- **Temporal scope:** Using 2010–2012 data limits generalisability to today's retail (e-commerce integration, supply-chain shifts, changing consumer behaviour). Subsequent industry changes may alter key drivers and model effectiveness.
- **Train/test windows**: An 80-week training window and 8-week test horizon may miss longer-term structural shifts and alternative planning horizons used operationally.

Methodological limitations

- **Aggregation level**: Chain-level results mask store/department heterogeneity (seasonality, promotion sensitivity, trends) that could change method rankings—though M5 evidence often shows larger ML gains at disaggregate levels.
- **External variables:** To isolate temporal signal, we omit promotions, weather, fuel, and macro indicators. This controlled setup may understate gains from multivariate models and limiting applicability to data-rich modern retail.
- **Hyperparameters:** Fixed ML hyperparameters prioritise comparability over per-model optimisation; deeper tuning could improve individual models but risks overfitting.

Algorithmic scope

- **Tree-based focus:** Emphasis on tree ensembles (for compute and consistency) excludes deep learning (e.g., LSTM, CNN) that can perform well with sufficient data; however, Hall & Rasheed (2025) report tree methods often outperform DL with far better efficiency.

External validity

- **Retail specificity:** Findings may not generalize beyond retail contexts with similar seasonal patterns, promotional structures, and demand characteristics; other sectors may yield different method rankings.
- **Implementation assumptions:** Economic impact estimates assume particular cost structures and inventory policies; variation across firms limits direct transferability.

5.4 Recommendations for Practice and Research

Research Priorities

Future research could explore several promising methodological extensions. Hierarchical forecasting approaches that model chain, store, and department-level demand simultaneously could provide improved accuracy through information sharing across aggregation levels whilst supporting decision-making at multiple organisational levels.

Deep learning approaches, particularly Long Short-Term Memory (LSTM) networks and attention mechanisms, could potentially capture complex temporal dependencies that exceed the capabilities of traditional machine learning algorithms. However, such methods would require careful evaluation of performance improvements relative to increased complexity and data requirements.

The incorporation of external variables through ARIMAX, Vector Autoregression (VAR), or multivariate machine learning approaches represents another promising research direction. The demonstrated importance of temporal features suggests that combining these patterns with promotional, economic, and competitive variables could yield substantial accuracy improvements.

Hierarchical and cross-sectional forecasting approaches represent another critical research priority, enabling organisations to optimise forecasting accuracy across multiple decision-making levels whilst maintaining consistency between strategic and operational planning processes.

The development of adaptive forecasting systems that can automatically adjust model selection and feature importance based on evolving business conditions represents a promising direction for bridging academic research and operational requirements.

The substantial performance advantages observed across the machine learning family suggest that the field should prioritize maximizing machine learning effectiveness rather than preserving classical method relevance, though balanced approaches that consider organizational constraints and business requirements remain essential for practical success.

Immediate Implementation Recommendations:

1. **Start with Prophet:** Begin with Prophet implementation to capture immediate 13.0% accuracy improvements while building analytical capabilities for more sophisticated approaches.
2. **Develop Feature Engineering Capabilities:** Invest in automated temporal feature engineering pipelines focusing on lag variables (particularly lag_52), rolling statistics, and seasonal indicators as reusable analytical assets.

3. **Establish Robust Validation:** Implement time series cross-validation frameworks with multiple temporal splits to ensure reliable performance assessment and avoid overfitting artifacts.
4. **Plan Progressive Enhancement:** Design implementation roadmaps progressing from Prophet (13.0% improvement) to XGBoost (30.4% improvement) to Random Forest (38.3% improvement) as analytical sophistication increases.
5. **Maintain Classical Benchmarks**: Retain simple classical methods (particularly Seasonal Naïve) as performance monitoring tools and sanity checks rather than primary forecasting engines.

Strategic Considerations: For resource-constrained organizations, classical methods may retain value in contexts requiring maximum transparency or minimal computational resources. However, for most retail forecasting applications where accuracy improvements translate to significant business value, machine learning adoption appears increasingly justified. The key insight is matching method sophistication to organizational capabilities while maintaining operational reliability and business insight generation that decision-makers require.

6.CONCLUSION

In conclusion, this study provides a comprehensive and methodologically robust comparison of classical time series models and machine learning approaches for weekly retail sales forecasting at a chain level. The empirical evidence robustly demonstrates that machine learning models—Random Forest, XGBoost, and Prophet—substantially outperform classical benchmarks including Seasonal Naïve, SARIMA, and ETS. This superior performance remains consistent across multiple temporal validations, underscoring the systematic advantage of machine learning models in capturing the complex, nonlinear dynamics inherent in retail sales data.

The study's findings challenge longstanding assumptions in forecasting literature regarding the adequacy of linear and classical methods for modern retail environments characterized by irregular promotional effects, complex seasonality, and diverse external influences. Importantly, the provision of identical feature sets to all models ensures that these performance differences reflect genuine algorithmic capabilities rather than disparities in data access, addressing critical methodological limitations identified in recent comparative research. The systematic use of Monte Carlo cross-validation strengthens the reliability of the results and highlights the importance of rigorous temporal validation in retail forecasting research.

From a practical standpoint, even modest gains in Mean Absolute Percentage Error translate to substantial economic value at scale—through enhanced inventory management, reduced stockouts, and improved promotional planning. The dominant predictive value of lagged sales and seasonal indicators, which surpassed initial expectations regarding the relative importance of short-term versus long-term temporal features, illustrates the critical role of feature engineering in enabling machine learning models to uncover fundamental patterns that classical methods model explicitly but less flexibly. While Prophet offers a balance between interpretability and accuracy, tree-based ensembles excel in predictive power and consistency, supporting their broader adoption in retail forecasting applications. Nonetheless, classical methods may retain relevance in scenarios demanding maximum interpretability or where computational simplicity is paramount, especially in less complex or data-sparse environments.

This study's rigorous design and empirical rigour have yielded robust and reproducible findings, and I am pleased that the results consistently affirmed machine learning superiority under fair evaluation conditions. Nevertheless, certain constraints—such as reliance on chain-level data aggregation and fixed hyperparameter configurations—may have limited the generalisability of specific performance estimates. In future work, I would explore more extensive hyperparameter optimisation, investigate hierarchical forecasting frameworks at both store and product levels, and integrate richer external datasets (for example, promotional calendars and competitor pricing) to further enhance model adaptability.

Addressing noted research gaps, this study emphasises the need for ongoing development of systematic feature engineering frameworks, expanded evaluation protocols that incorporate external contextual variables, and improved interpretability techniques tailored for time series forecasting. By bridging theoretical understanding and practical deployment, this dissertation lays a foundational step toward smarter, data-driven retail forecasting systems that can dynamically adjust to evolving market conditions and complex consumer behaviours.

7.REFERENCES

Ahmadov, A. and Helo, P., 2023. Deep learning-based approach for forecasting intermittent online sales. Discover Artificial Intelligence, 3(1), p.45.

Bergmeir, C. and Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, pp.192-213.

Brykin, D., 2024. Sales Forecasting Models: Comparison between ARIMA, LSTM and Prophet. Journal of Computer Science, 20(10), pp.1222-1230.

Hall, T. and Rasheed, K., 2025. A survey of machine learning methods for time series prediction. Applied Sciences, 15(11), p.5957.

Hasan, M. R., Kabir, M. A., Shuvro, R. A., & Das, P. (2022). A Comparative Study on Forecasting of Retail Sales. arXiv:2203.06848 [cs.LG].

Hobor, L., Brcic, M., Polutnik, L., & Kapetanovic, A. (2025). Comparative Analysis of Modern Machine Learning Models for Retail Sales Forecasting. arXiv:2506.05941 [cs.LG].

Hyndman, R.J. and Athanasopoulos, G., 2018. Forecasting: Principles and practice (2nd ed.) [Online]. Melbourne: OTexts. Available from: https://otexts.com/fpp2/ [Accessed 1 July 2025].

İnce, M.N. and Taşdemir, Ç., 2024. Forecasting retail sales for furniture and furnishing items through the employment of multiple linear regression and Holt–Winters models. Systems, 12(6), p.219.

Jahin, M.A., Shahriar, A., Al Amin, M., et al., 2024. MCDFN: Supply chain demand forecasting via an explainable multi-channel data fusion network model. arXiv preprint arXiv:2405.15598.

Kontopoulou, V.I., Panagopoulos, A.D., Kakkos, I. and Matsopoulos, G.K., 2023. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. Future Internet, 15(8), p.255.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V., 2018. Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE, 13(3), p.e0194889.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V., 2022. M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting, 38(4), pp.1346-1364.

McKinsey, 2024. AI-driven operations forecasting in data-light environments. McKinsey & Company [Online]. Available from: https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments [Accessed 1 September 2025].

Mitra, P., Rabenoro, O.R. and Weber, J., 2022. Combining statistical and machine learning models for network traffic forecasting. IEEE Transactions on Network and Service Management, 19(1), pp.678–691.

Nasseri, M., Falatouri, T., Brandtner, P. and Darbanian, F., 2023. Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning. Applied Sciences, 13(19), p.11112.

Oreshkin, B.N., Carpov, D., Chapados, N. and Bengio, Y., 2020. N-BEATS: Neural basis expansion analysis for interpretable time series. In: International Conference on Learning Representations (ICLR).

Petropoulos, F., Grushka-Cockayne, Y. and Siemsen, E., 2021. Wielding Occam's razor: Fast and frugal retail forecasting. arXiv preprint arXiv:2104.12345.

Rabenoro, T., Bien-Barkowska, K., Csanadi, A. and Pálvölgyi, Á., 2022. Retail demand forecasting: A comparative study of machine learning approaches. IEEE Access, 10, pp.48251-48267.

Smyl, S., 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting, 36(1), pp.75-85.

Statista, 2025. Retail trade in the UK. Statista [Online]. Available from: https://www.statista.com/topics/8671/retail-trade-in-the-uk/ [Accessed 1 September 2025].

Taylor, S.J. and Letham, B., 2018. Forecasting at scale. The American Statistician, 72(1), pp.37-45.

Xian, X., Wang, L., Wu, X., Tang, X., Zhai, X., Yu, R., Qu, L. and Ye, M., 2023. Comparison of SARIMA model, Holt-Winters model and ETS model in predicting the incidence of foodborne disease. BMC Infectious Diseases, 23(1), p.803.

Žunić, E., Korjenić, K., Hodžić, K., & Đonko, D. (2020). Application of Facebook's Prophet algorithm for successful sales forecasting based on real-world data. IJCSIT, 12(2), 23–29.