



SCHOOL OF SOCIAL SCIENCES

SUPPLY CHAIN MANAGEMENT
(SCM)

POSTGRADUATE DISSERTATION

**Unveiling Trends and Patterns in Walmart Sales:
A Time Series Analysis Approach**

Author

Pavlos Miaoulis

Supervisor

Nikolaos Thomaidis

Patras, January 2025

Theses/Dissertations remain the intellectual property of students (“authors/creators”), but in the context of open access policy they grant to the HOU a non-exclusive license to use the right of reproduction, customization, public lending, presentation to an audience and digital dissemination thereof internationally, in electronic form and by any means for teaching and research purposes, for no fee and throughout the duration of intellectual property rights. Free access to the full text for studying and reading does not in any way mean that the author/creator shall allocate his/her intellectual property rights, nor shall he/she allow the reproduction, republication, copy, storage, sale, commercial use, transmission, distribution, publication, execution, downloading, uploading, translating, modifying in any way, of any part or summary of the dissertation, without the explicit prior written consent of the author/creator. Creators retain all their moral and property rights.



Unveiling Trends and Patterns in Walmart Sales: A Time Series Analysis Approach

Author

Pavlos Miaoulis

Supervising Committee

Supervisor:

Nikolaos Thomaidis

Associate Professor

Department of Financial &

Management Engineering

University of the Aegean

Co-Supervisor:

Antonios Demos

Professor

Department of International &

European Economic Studies

Athens University of Economics and

Business

Patras, January 2025

"To my beloved wife and lifetime companion."

Abstract

Retail demand forecasting is a key component of modern supply chain management that allows businesses to optimize their inventory levels thus reducing costs and meeting consumer needs. This dissertation uses a rich data set from Walmart stores provided for the M5 Forecasting Competition to analyse retail sales patterns and drivers. Focusing on the three product categories aggregated to a weekly frequency, the study employs statistical tools and regression analysis to uncover trends, seasonality and the impact of exogenous variables such as pricing, holidays-events and promotions on demand.

The objective of this study is to demonstrate that a detailed exploratory analysis and a deep understanding of time series dynamics are crucial to obtain accurate demand forecasts. By applying a combination of regression and exponential smoothing methods and assessing their performance through statistical accuracy metrics, the study attempts to identify the best predictive model for each category and highlight the importance of a data-driven approach in demand forecasting.

Keywords: Demand forecasting, retail sales, time series analysis, regression analysis

Περίληψη

Η πρόβλεψη της ζήτησης στο λιανικό εμπόριο διαδραματίζει καθοριστικό ρόλο στη σύγχρονη διαχείριση της εφοδιαστικής αλυσίδας, επιτρέποντας στις επιχειρήσεις να βελτιστοποιούν τα αποθέματα, να μειώνουν το κόστος και να ανταποκρίνονται στις ανάγκες των καταναλωτών. Η παρούσα εργασία διερευνά τα πρότυπα και τους παράγοντες που επηρεάζουν τις λιανικές πωλήσεις, αξιοποιώντας ένα ολοκληρωμένο σύνολο δεδομένων από καταστήματα της Walmart, το οποίο παρασχέθηκε στο πλαίσιο του Διαγωνισμού Πρόβλεψης M5. Εστιάζοντας στις τρεις προϊόντικές κατηγορίες, σε εβδομαδιαία συχνότητα, η μελέτη χρησιμοποιεί στατιστικά εργαλεία και ανάλυση παλινδρόμησης για να αποκαλύψει τάσεις, εποχικότητα και την επίδραση εξωγενών παραγόντων, όπως η τιμή, γιορτές-αργίες και οι προωθητικές ενέργειες, στη ζήτηση.

Στόχος της μελέτης είναι να καταδείξει ότι μια ενδελεχής ανάλυση και μια σε βάθος κατανόηση της δυναμικής των χρονοσειρών, αποτελούν βασικούς παράγοντες για την εξαγωγή προβλέψεων ζήτησης. Εφαρμόζοντας έναν συνδυασμό μεθόδων παλινδρόμησης και εκθετικής εξομάλυνσης και αξιολογώντας την απόδοσή τους μέσω στατιστικών ελέγχων ακρίβειας, η μελέτη επιχειρεί να εντοπίσει το βέλτιστο μοντέλο πρόβλεψης για κάθε κατηγορία και να αναδείξει τη σημασία μιας δεδομενο-κεντρικής προσέγγισης στην πρόβλεψη ζήτησης.

Λέξεις-Κλειδιά: Πρόβλεψη ζήτησης, λιανικές πωλήσεις, ανάλυση χρονοσειρών, ανάλυση παλινδρόμησης

Contents

Abstract	v
Contents	vii
Figures	viii
Tables	ix
Abbreviations	x
1. Introduction	1
1.1. Problem Statement	1
1.2. Research Scope	1
1.3. Research Outline	3
2. Research Methodology	4
2.1. Data Modeling Framework	4
2.2. Data Exploration Methods	5
2.2.1. Preliminary Analysis	5
2.2.2. Autocorrelation Functions	6
2.2.3. STL Decomposition	7
2.2.4. Time Series Stationarity	7
2.3. Regression Analysis	10
2.3.1. Linear Regression	10
2.3.2. Evaluation of Linear Regression	12
2.4. Exponential Smoothing	17
2.4.1. Simple Exponential Smoothing (SES)	17
2.4.2. Holt-Winters Method	18
2.5. Accuracy Evaluation	19

3. Data Overview and Preprocessing	21
3.1. Overview of the Dataset	21
3.1.1. The M5 Competition	21
3.1.2. Raw Data Overview	22
3.2. Data Preprocessing	23
3.2.1. Combining the Data	23
3.2.2. Data Aggregation	26
3.2.3. Feature Engineering	26
4. Exploratory Data Analysis (EDA)	28
4.1. Descriptive Statistics	28
4.2. Distributional Analysis	29
4.3. Categorical Exploration	32
4.4. Cross-Sectional Analysis	34
4.5. Key Takeaways	35
5. Time Series Analysis	36
5.1. Time Series Overview	36
5.1.1. FOODS	37
5.1.2. HOBBIES	37
5.1.3. HOUSEHOLD	38
5.2. Seasonality Visualization	38
5.3. Time Series Decomposition	41
5.4. Autocorrelation and Lags Analysis	44
5.5. Stationarity Tests	49
5.6. Categorical Variables Over Time	50
5.7. Key Takeaways	52
6. Regression Analysis	53
6.1. Modeling Process	53
6.2. Regression Analysis - FOODS	54
6.2.1. Initial Regression Model	54
6.2.2. Residuals Diagnostics	54
6.2.3. Refined Regression Model	56
6.2.4. Residual Diagnostics for the Refined Model	56
6.2.5. Fit Assessment	58
6.3. Regression Analysis - HOBBIES	60
6.3.1. Initial Regression Model	60

6.3.2.	Residual Diagnostics	61
6.3.3.	Refined Regression Model	62
6.3.4.	Residual Diagnostics for the Refined Model	63
6.3.5.	Fit Assessment	64
6.4.	Regression Analysis - HOUSEHOLD	66
6.4.1.	Initial Regression Model	66
6.4.2.	Residual Diagnostics	66
6.4.3.	Refined Regression Model	68
6.4.4.	Residual Diagnostics for the Refined Model	68
6.4.5.	Fit Assessment	70
7.	Out-of-Sample Model Evaluation	73
7.1.	Dataset Preprocessing	73
7.2.	Modeling Process	74
7.2.1.	FOODS Models	75
7.2.2.	HOBBIES Models	77
7.2.3.	HOUSEHOLD Models	80
7.3.	Accuracy Evaluation	83
8.	Main Findings and Conclusion	85
8.1.	Research Questions and Findings	85
8.2.	Conclusion	86
8.3.	Recommendations for Future Research	87
	References	88
	Appendices	91
A.	Appendix A	91
A.1.	Supplementary Analysis Tables	91
A.2.	Supplementary Modeling Tables	96
B.	Appendix B	100
B.1.	R Code for Data Preprocessing	100
B.2.	R Code for Data Modeling	102

Figures

2.1. Modelling framework	4
3.1. Hierarchies of the M5 competition data	23
3.2. Entity-relationship diagram of joins used for creating the unified dataset	24
4.1. Distribution plots of the sales variable	29
4.2. Violin plots of weekly sales by category	30
4.3. Distribution plots of the price variable	31
4.4. Violin plots of weekly mean price by category	31
4.5. Total units sold and total distinct products sold by category	32
4.6. Boxplots of weekly sales by number of SNAP days in a week	33
4.7. Boxplots of weekly sales by event type	33
4.8. Scatter plots of weekly sales vs weekly mean price	34
4.9. Correlation of numerical variables by category	35
5.1. Time series plots of weekly sales over time	36
5.2. Seasonal plots of monthly sales	39
5.3. Seasonal plots of weekly sales	40
5.4. Boxplots of sales by week-of-month	41
5.5. Time series components of the FOODS category	42
5.6. Time series components of the HOBBIES category	43
5.7. Time series components of the HOUSEHOLD category	44
5.8. Autocorrelation plots of sales by category	45
5.9. Partial autocorrelation plots of sales by category	46
5.10. Lag plots of weekly FOODS sales	47
5.11. Lag plots of weekly HOBBIES sales	48
5.12. Lag plots of weekly HOUSEHOLD sales	49
5.13. Weekly sales with points color-coded by number of SNAP days in a week	51
5.14. Impact of the last week of month on sales over time	52

6.1. Residuals diagnostic plots of the initial FOODS regression model fit	55
6.2. Residuals diagnostics plots of the refined FOODS regression model fit	57
6.3. Plot of the refined FOODS model fit vs original values	60
6.4. Residuals diagnostic plots of the initial HOBBIES regression model fit	61
6.5. Residuals diagnostic plots of the refined HOBBIES regression model fit	63
6.6. Plot of the refined HOBBIES model fit vs original values	66
6.7. Residuals diagnostic plots of the initial HOUSEHOLD regression model fit . . .	67
6.8. Residuals diagnostic plots of the refined HOUSEHOLD regression model fit . .	69
6.9. Plot of the refined HOUSEHOLD model fit vs original values	72
7.1. Training - Test split process	73
7.2. Plot of the FOODS models fit on the test set	76
7.3. Plot of the FOODS models residuals from the test set	77
7.4. Plot of the HOBBIES models fit on the test set	79
7.5. Plot of the HOBBIES models residuals from the test set	80
7.6. Plot of the HOUSEHOLD models fit on the test set	81
7.7. Plot of the HOUSEHOLD models residuals from the test set	82

Tables

3.1. Sample rows from the joined dataset	25
4.1. Summary statistics for the continuous variables	28
5.1. Results of stationarity tests	50
6.1. Tests results on the residuals of the initial FOODS model fit	56
6.2. Tests results on the residuals of the refined FOODS model fit	58
6.3. Summary of the refined regression model for the FOODS time series	59
6.4. Tests results on the residuals of the initial HOBBIES model fit	62
6.5. Tests results on the residuals of the refined HOBBIES model fit	64
6.6. Summary of the refined regression model for the HOBBIES time series	65
6.7. Tests results on the residuals of the initial HOUSEHOLD model fit	68
6.8. Tests results on the residuals of the refined HOUSEHOLD model fit	70
6.9. Summary of the refined regression model for the HOUSEHOLD time series	71
7.1. Accuracy results for the in-sample fit of the models	83

Abbreviations

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller test
CA	State of California
CV	Coefficient of Variation
EDA	Exploratory Data Analysis
GVIF	Generalized Variance Inflation Factor
HW	Holt-Winters method
IQR	Interquartile Range
KPSS	Kwiatkowski–Phillips–Schmidt–Shin test
LM	Linear Regression Model
LOESS	Locally Estimated Scatterplot Smoothing
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
OLS	Ordinary Least Squares
PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SD	Standard Deviation
SES	Simple Exponential Smoothing method
SNAP	Supplemental Nutrition Assistance Program
STL	Seasonal & Trend Decomposition using LOESS
TX	State of Texas
USA	United States of America
VIF	Variance Inflation Factor
WI	State of Wisconsin

1. Introduction

1.1. Problem Statement

Demand Forecasting is an essential challenge in the modern retail supply chain. As a result of growing competition in worldwide markets, the ability to accurately forecast demand has become an essential element of strategic decision making. Understanding consumer behavior and analyzing demand patterns is essential for organizations in order to optimize their inventory levels, align production schedules and plan promotions effectively while at the same time enabling the minimization of costs, reducing waste and improving service.

Retail sales data contain both deterministic and stochastic elements. By breaking time series into their component parts, we can generate useful insights. Trends and seasonality provide a predictable structure, capturing long-term growth and recurring patterns, while short-term variations driven by factors like promotions, pricing changes and external influences add volatility and localized fluctuations. Together, these characteristics make retail time series both rich in information and challenging to model effectively.

Additional complexity arises from the need to distinguish structured time series data from purely random processes, such as random walk or white noise. Misinterpreting stochastic elements as randomness can lead to inaccurate predictions and bad decisions, particularly when external factors such as holidays, promotions and pricing strategies introduce further variability into the data. Identifying and isolating these effects requires robust analytical methods capable of separating signal from noise.

1.2. Research Scope

This study aims to address the above-mentioned challenges of retail demand forecasting through the analysis of data from the M5 Forecasting Competition, an open forecasting contest specifically designed for predicting retail sales. The fact that the M5 datasets contain real-life sales data

makes them an ideal resource for investigating demand patterns and identifying key influential variables, replicating real-world scenarios.

The main research objective of this project is to investigate the demand patterns for various product categories, focusing on trends, seasonal behaviors and recurring sales dynamics. Secondly, it aims to analyze the impact of factors such as pricing, promotions, calendar-specific variables and other temporal influences on sales performance, providing insights into how these variables shape consumer behavior and retail demand.

The study addresses the following key research questions:

1. What are the demand patterns across product categories? How do trends, seasonality and recurring sales dynamics manifest in retail sales time series for different product categories?
2. What are the critical factors influencing retail sales behavior? How do pricing strategies, promotions, calendar events and temporal factors impact sales performance and consumer behavior?
3. How forecastable are retail sales time series? What is the typical accuracy rate of weekly sales forecasts and how does this accuracy vary across product categories?
4. How do deterministic and stochastic components interact in retail sales time series? How can the combination of structured patterns and short-term variations enhance the understanding of sales dynamics and improve forecasting methodologies?

This study seeks to answer a set of questions aimed at uncovering meaningful patterns and relationships within retail sales data that could help to substantially improve the accuracy of demand predictions. Taking the perspective of a demand planner at Walmart, this project focuses on the critical role of forecasting in optimizing supply chain operations. Effective forecasting is essential for enabling the alignment of inventory levels with actual demand in order to minimize costs and avoid stockouts. This also leads to better alignment between manufacturing timelines and sales, leading to reduced wastes and more sustainable practices. Additionally, accurate forecasts improve the organization's capacity to adapt to changing market conditions and consumer needs. By differentiating deterministic patterns from stochastic noise and evaluating the forecastability of retail sales time series, this research provides a practical framework for leveraging data to improve planning and drive supply chain optimization within a large retail organization.

1.3. Research Outline

The dissertation is structured as follows:

- **Chapter 1** is the introductory part of the study where the scope and research questions are presented.
- **Chapter 2** provides the theoretical framework of the model building process that the study follows, together with an overview of the statistical methods used throughout the process.
- **Chapter 3** details the data sources, preprocessing steps and feature engineering processes applied to prepare the dataset for analysis.
- **Chapter 4** presents exploratory data analysis, focusing on uncovering the main characteristics of the given variables, identifying and quantifying the relationships between them.
- **Chapter 5** focuses on time series analysis to identify temporal behavior, apply decomposition techniques, stationarity assessment and evaluate the influence of variables over time.
- **Chapter 6** applies regression modeling to quantify the relationships between sales and explanatory variables, with a focus on deterministic and short memory effects.
- **Chapter 7** applies the insights gathered from previous chapters into building appropriate predictive models and then evaluates the predictive accuracy of these models, through out-of-sample testing to find the best performing method.
- **Chapter 8** summarizes and discusses the key findings as well as outlines areas for future research.

2. Research Methodology

2.1. Data Modeling Framework

The data modelling framework used in this study as proposed by Hyndman & Athanasopoulos (2021), follows a structured approach to produce accurate results. The stages of this modeling framework are:

1. **Data Preparation:** The initial step is to organize the data in an appropriate format for analysis. The tasks include merging data sources, loading the data and applying transformations. Proper preparation of the data ensures that subsequent exploratory analysis and modeling are effective.

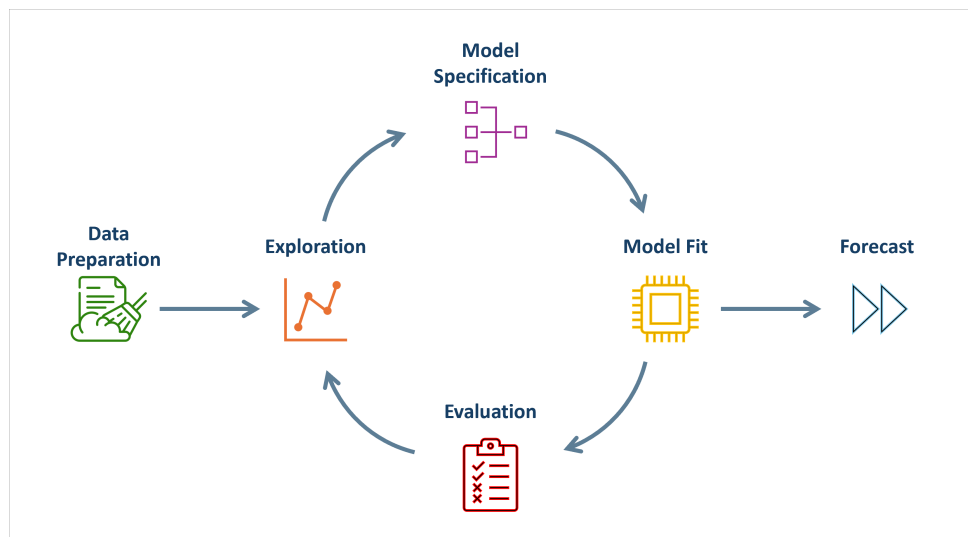


Figure 2.1.: Modelling framework

2. **Exploration:** Visual and statistical exploration helps to understand the data and identify patterns. The insights obtained from this stage will guide the selection of appropriate modeling techniques.

3. **Model Specification:** Based on insights from the previous step, a suitable model is chosen. This involves defining the variables that will be used so that the time series attributes are accurately modeled.
4. **Model Fit:** With the model specified, the next step is to fit the data to create predictions that accurately represent the already identified underlying patterns.
5. **Evaluation:** After fitting the model to the data, diagnostics tests are applied on the residuals to evaluate its performance. The evaluation checks for any pattern or statistical behavior in the residuals that would signify any model inadequacy.
6. **Forecasting:** Once the model has been validated, it is used to generate forecasts for future periods. These forecasts provide insights into expected future trends and are utilized for informed decision-making.

2.2. Data Exploration Methods¹

2.2.1. Preliminary Analysis

For the first part of the data exploration, descriptive statistics are used to summarize variables' key features. Measures of central tendency (mean, median) and dispersion (standard deviation, range) give a first idea of the variability and distribution of continuous variables. Moreover, advanced visual aids are used to further examine as well as interpret the data. Different types of plots are used like Histograms, Boxplots and Violin plots to explore the distribution and variability of numerical data. In addition, Bar plots compare categorical variables across different groups and Scatter plots explore relationships and interactions between variables. To check variations and get a dynamic view of changes over time, Time Series plots are used. These plots are particularly useful for the identification of trends and seasonality as well.

¹All statistical analysis, data visualization and data modeling in this research were conducted with the R software version 4.2.2 (R Core Team, 2024) and the following main R packages: gtsummary v. 2.0.4 (Sjoberg et al., 2024), fable v. 0.4.1 (O'Hara-Wild et al., 2024a), feasts v. 0.4.1 (O'Hara-Wild et al., 2024b), kableExtra v. 1.4.0 (Zhu, 2024), lmttest v. 0.9-40 (Hothorn et al., 2022), tidyverse v. 2.0.0 (Wickham, 2023), tseries v.0.10-58 (Trapletti & Hornik, 2024), tsibble v.1.1.5 (Wang et al., 2024). This document was prepared and rendered in PDF format using the open-source publishing system Quarto v. 1.5.57 (Allaire et al., 2024).

Pearson's Coefficient

Pearson's correlation coefficient (r) is a measure of the degree of the linear relationship between the numeric variables in the dataset. It indicates whether a change in one variable is related to a change in the other. The r value can range from -1 (perfect negative relationship) to 1 (perfect positive relationship) with 0 indicating no linear relationship. The formula for Pearson's correlation coefficient is given as follows:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2.1)$$

here X and Y are the individual data points of variables X and Y respectively. A correlation matrix that visualizes the Pearson coefficients helps in understanding the relation between the variables intuitively. In a correlation matrix, the correlation coefficient for each pair of variables is shown as a cell. Color palettes are generally used to render shades for positive and negative values, highlighting type of the dependency or multicollinearity between a datasets variables.

2.2.2. Autocorrelation Functions

ACF

The Autocorrelation Function (ACF) is a method used to measure the degree of linear relationship between a time series and lagged values of the time series. In this analysis, ACF is applied prior to modelling as a way to check dependency and spot repetitive patterns. Post-modeling, the ACF of residuals is used as a diagnostic tool to check for remaining autocorrelation, ensuring that the model adequately captures the underlying structure of the data.

The empirical ACF for a time series X_t at lag k is computed as:

$$r_k = \frac{\sum_{t=k+1}^T (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^T (Y_t - \bar{Y})^2} \quad (2.2)$$

where T is the length of the time series. Values near ± 1 indicate strong autocorrelation, while values near 0 suggest minimal dependency (Hyndman & Athanasopoulos, 2021).

Plotting the ACF results is a common practice, commonly referred to as the ACF plot. By visualizing the autocorrelation at multiple lags, we can spot patterns like peaks at regular intervals,

indication of seasonality or decaying behavior where the gradual fading or persistence of the peaks suggests non-stationarity and the presence of trend.

PACF

The Partial Autocorrelation Function (PACF) is another autocorrelation testing tool checking the correlation between two observations after accounting for the contributions of intermediate lags. Partial autocorrelation coefficient ϕ_{kk} , as defined by Box et al. (2016), measures the correlation between Y_t and Y_{t-k} while removing the linear influence of all intermediate lags $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$ in the time series. This modification guarantees ϕ_{kk} only informs on Y_t and Y_{t-k} and not on the shorter time lags, thereby excluding their influence. In this study the PACF plot is used to identify the lags at which significant direct relationships persist in order to incorporate these lags in the modeling process to adjust for autocorrelation in residuals and improve the validity of the models.

2.2.3. STL Decomposition

Seasonal and Trend decomposition with LOESS (STL) refers to a method for decomposing time series into seasonal, trend and remainder components, introduced by Cleveland et al. (1990). For this, STL uses a non-parametric regression method that fits a smooth curve to the data points called Locally Estimated Scatterplot Smoothing (LOESS).

Starting from the Seasonal component, STL breaks the data into seasonal subseries (e.g., all Januaries, all Februaries etc) and applies LOESS to smooth the values across years. By combining all the seasonal curves we get back the resulted overall seasonal component. Next, by extracting this seasonal component from the original data (seasonally adjusted series), STL estimates the trend component by re-applying LOESS to fit a smooth curve that will catch the overall direction of the series. These two steps (seasonal & trend smoothing) are repeated iteratively until the components stabilize. Finally, the remainder component is calculated by subtracting both the seasonal and trend components from the original time series. This last random component is the feature of the data that cannot be explained by the seasonal or trend components.

2.2.4. Time Series Stationarity

Stationarity is a key concept in time series analysis. A time series is considered stationary if both its mean and variance do not change over time. Time series models must be stable over time in

order to ensure that relationships between values remain stable. We need to analyze stationarity so we can better understand the underlying behavior of a time series, choose appropriate methods or assess the performance of a model through its residuals.

ADF test

The augmented Dickey-Fuller (ADF) test is a statistical test used to locate the presence of a unit root in a univariate time series. A unit root is a characteristic of a time series where shocks to the series cause it to lack a fixed mean and exhibit a stochastic trend. The original Dickey-Fuller test fits a regression with a constant term, a trend term and a lagged value of the series to test for a unit root (Dickey & Fuller, 1981). The ADF test extends the original test by including these lagged differences of the series in the regression, which allows it to become more robust. The regression is formulated like:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^p \delta_i \Delta Y_{t-i} + \varepsilon_t \quad (2.3)$$

where:

- $\Delta Y_t = Y_t - Y_{t-1}$: The first difference of the series
- α : Constant term
- β : Trend component
- γ : Coefficient that determines the presence of a unit root (0/1)
- $\sum_{i=1}^p \delta_i \Delta Y_{t-i}$: Sum of lagged differences to account for autocorrelation
- ε_t : The error term

The test involves calculating the test statistic and comparing it to critical values to determine whether to reject the null hypothesis. The null hypothesis of the test is that the time series has a unit root ($H_0 : \gamma = 0$) and the alternative is that the time series is stationary ($H_1 : \gamma < 0$). The test statistic is calculated as:

$$t_\gamma = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \quad (2.4)$$

Based on the test statistic, the p-value is computed and if it is found below the chosen 0.05 significance level, the null hypothesis of a unit root is rejected, indicating stationarity (Said & Dickey, 1984).

KPSS test

The KPSS test is another method used for examining the trend stationarity of a time series. The KPSS test starts out with the assumption of stationarity as opposed to the ADF which checks for a unit root (non-stationarity). This different logic means it can be used as a complimentary to the ADF test, allowing us to look at time series stationarity from different perspectives. The initial stage involves applying a linear regression model to separate the time series into three components: a random walk, a deterministic trend and a stationary error component. Then, it tests the residuals for stationarity. The regression used is of the type:

$$X_t = r_t + \beta_t + \epsilon_t \quad (2.5)$$

where:

- r_t is the random walk component
- β_t is the deterministic trend component
- ϵ_t is the error (residuals)

After performing the regression, the cumulative sum of residuals is calculated ($S_t = \sum_{i=1}^t \epsilon_i$), which represents the deviations from the deterministic trend over time. The test statistic is then calculated by evaluating the variance of S_t , normalized by the variance of the residuals (σ^2) like:

$$\text{KPSS} = \frac{1}{T^2 \sigma^2} \sum_{t=1}^T S_t^2 \quad (2.6)$$

where:

- T is the number of observations
- σ^2 is the variance of the residuals
- S_t is the cumulative sum of residuals

The null hypothesis (H_0) of the KPSS is that the time series is trend-stationary around a deterministic trend and the alternative hypothesis (H_1) is that there is a unit root in the time series making data to wander like a random walk.

The test statistic calculated from the cumulative sum of residuals is compared to pre-tabulated critical values and if the test statistic exceeds the critical value at the chosen significance level

(5%), we reject the null hypothesis, indicating evidence of non-stationarity. Again, the p-value is computed to quantify the probability of observing the test statistic under the null hypothesis. A small p-value ($p < 0.05$) provides strong evidence against H_0 , while a large p-value suggests insufficient evidence to reject H_0 , supporting trend-stationarity (Kwiatkowski et al., 1992).

2.3. Regression Analysis

2.3.1. Linear Regression

Linear Regression Model

Linear regression is commonly used in statistics and econometrics to analyze and model the relationship between one or more independent variables and a dependent variable. Its purpose is to explain how changes in the independent variables influence the dependent variable, allowing us to uncover patterns, make predictions and gain meaningful insights from data. There are mainly two different types of linear regression, Univariate Linear Regression (or Simple Linear Regression) and Multivariate Linear Regression (or Multiple Linear Regression).

In univariate linear regression, we focus on modeling the relationship between one explanatory variable (X) and one dependent variable (Y). This kind of model is expressed as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (2.7)$$

where:

- β_0 represents the intercept (the value of Y when $X = 0$),
- β_1 represents the slope (the change in Y for a one-unit change in X),
- ε is the error term capturing the deviation of observed values from the predicted values.

Here, the relationship between the independent variable (X) and the dependent variable (Y) is explored by estimating the line that best fits the data points. The slope (β_1) quantifies how much Y changes for a one-unit change in X , while the intercept (β_0) represents the value of Y when $X = 0$. Finally, the error term (ε) captures deviations that the linear relationship cannot explain.

In multivariate linear regression we are able to explore how more than one independent variables are influencing the dependent variable. The relationship between several independent variables

(X_1, X_2, \dots, X_p) and a dependent variable (Y) is explored by estimating their combined effect in a model expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (2.8)$$

where:

- X_1, X_2, \dots, X_p represent the independent variables and
- $\beta_1, \beta_2, \dots, \beta_p$ are their corresponding coefficients.

Each coefficient β_i quantifies the change in Y associated with a one-unit change in the corresponding independent variable X_i , while holding all other variables constant. The intercept β_0 represents the predicted value of Y when all independent variables are zero and the error term ε accounts for the variability in Y not explained by the model. This approach allows for a comprehensive analysis of how multiple factors jointly influence the dependent variable.

In this study we will also exploit categorical variables (dummy variables) in order to incorporate to the models, information about seasonal periods or the presence of other influential factors like promotions and holidays.

Ordinary Least Squares method

Ordinary Least Squares (OLS) is probably the most common method for estimating the coefficients of a linear regression model by providing a systematic approach to fit a line that best represents the relationship between independent variables (X) and a dependent variable (Y). This is achieved by minimizing the sum of squared residuals (SSE). Residuals (ε_i) are the differences between observed values (Y_i) and predicted values (\hat{Y}_i), calculated as:

$$\varepsilon_i = Y_i - \hat{Y}_i \quad (2.9)$$

The principle of Least Squares Estimation ensures that the best-fit line minimizes the sum of the squared differences between observed and predicted values. Squaring the residuals prevents positive and negative deviations from canceling each other out and assigns greater weight to larger deviations. By minimizing the SSE, OLS produces the line that provides the most accurate and unbiased linear relationship between X and Y (Hyndman & Athanasopoulos, 2021).

According to the Gauss-Markov theorem², the estimators obtained through this approach are efficient, having the smallest variance among all linear unbiased estimators. OLS minimizes the SSE, defined as:

$$\text{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2.10)$$

ensuring that larger errors are penalized more heavily by squaring the residuals and ensuring balanced consideration of positive and negative deviations. The regression coefficients $(\beta_0, \beta_1, \dots, \beta_p)$ are adjusted to minimize this sum, producing the best-fit line (Ord et al., 2017). OLS is highly effective because of its simplicity and interpretability, working well for both univariate and multivariate regressions with small and large datasets. There are certain assumptions required for Ordinary Least Squares (OLS) under the Gauss-Markov theorem, which ensure that OLS produces the Best Linear Unbiased Estimators (BLUE). The assumptions are linearity, unbiased residuals, constant variance of residuals (homoscedasticity), no autocorrelation and no perfect multicollinearity among the independent variables.

2.3.2. Evaluation of Linear Regression

Assessing Overall Fit

When we assess a linear regression model, we are often first interested to see how much of the variation in Y can be explained by X (goodness-of-fit). For this, the R^2 coefficient (of determination) is used which shows the share of total variance in Y being explained by the the model. That is, it ranges from 0 (no variance explained) to 1 (all variance explained). But, a high R^2 does not always imply a good fit of the model. The model may be too complex which may lead to over-fit or that the OLS assumptions are violated, making the R^2 value not reliable. To fix this issue, we use the *Adjusted R^2* which adjusts the metric based on the number of predictors in the model. The *Adjusted R^2* penalizes irrelevant predictors, ensuring that the fit reflects meaningful contributions from the independent variables (Petropoulos et al., 2022).

One more important measure of overall model fit is the F-statistic which tests whether the model as a whole is significant. The F-statistic contrasts the total fit of the model against a null model which has no predictors. The p-value associated with the F-statistic tests the null hypothesis

²The Gauss-Markov theorem states that under certain assumptions, OLS estimators are the Best Linear Unbiased Estimators (BLUE). This means they have the smallest variance among all linear unbiased estimators, making them the most efficient for linear regression.

that all regression coefficients are equal to zero, which implies that the independent variables have no impact on Y . If p is smaller than 0.05, then at least one of the independent variables is contributing to Y . Beyond these metrics, residual plots can assist on assessing the fit. A good model will generate residuals that are random and have no obvious patterns. If patterns appear in the residuals, the model may have issues related to non-linearity, missing variables, or heteroscedasticity, suggesting the need for further investigation or adjustments.

Assessing Individual Variables

When looking at the overall fit of the model, one should also look at each variable independently, using the p-values, which denote whether a variable is found significant and whether it contributes to the overall fit. The p-value for each coefficient, tests the null hypothesis that the corresponding variable has no effect on the dependent variable ($\beta_i = 0$). A small p-value of less than 0.05 indicates the variable is probably not significant in explaining Y . If the p-values are large for certain variables, we can say that they do not contribute to the model meaningfully and therefore, can be dropped from the model. Moreover, the size and sign of the coefficients tell us how that particular variable is linked to Y . A positive coefficient indicates that an increase in the independent variable results in an increase in the dependent variable, while a negative coefficient indicates an inverse relationship.

Stationarity of Residuals

When the residuals of a regression model are stationary, this means that their temporal properties like their mean and variance remain constant over time. If the residuals are stationary, then the model has captured the trend and/or seasonality and/or any other structural component in the data, leaving only random noise. To evaluate the stationarity of the residuals, we will apply the ADF and KPSS tests together with the use of visuals, plotting the residuals over time to check for visible trends or other patterns. The residuals should present a stationary behavior in these plots resembling to a random noise time series fluctuating around zero and with constant variance. If any pattern or consistent behavior is visible from the plots, it means that there are probably issues with the model such as omitted variables or poor handling of trend or seasonality. By using visualizations along with the statistical tests we ensure that the stationarity assumption is fulfilled and whether the OLS model has captured the data structures properly.

Autocorrelation of Residuals

The autocorrelation assumption in OLS, states that the residuals should be independent of one another, meaning there should be no systematic relationship between residuals across time. Serial correlation, occurs when residuals from one time period are correlated with those from another, violating this assumption. If autocorrelation is present, it can lead to biased standard errors, affecting hypothesis testing and confidence intervals. To check for autocorrelation in residuals we will apply the Autocorrelation Function (ACF) and create the corresponding ACF plot to display the correlation of residuals with their lagged values across different time lags. In this plot we are looking for the bars of the first few lags to fall within the significance threshold (blue dashed lines). Systematic patterns of spikes outside the significance threshold would indicate a violation of the OLS assumption.

In addition to ACF plots, we will be applying a formal statistical test on the residuals, called the Ljung-Box test. This test checks for autocorrelation across multiple lags simultaneously. Its null hypothesis (H_0) assumes that the residuals are not autocorrelated (are independent), while the alternative hypothesis (H_1) suggests that autocorrelation exists at one or more lags (Ljung & Box, 1978). The Ljung-Box test statistic (Q) is calculated as:

$$Q^* = T(T + 2) \sum_{k=1}^h (T - k)^{-1} r_k^2 \quad (2.11)$$

where:

- T is the length of the time series
- h is the number of lags being tested
- r_k is the autocorrelation coefficient at lag k

The chi-squared (χ^2) distribution with h degrees of freedom is followed by the test statistic. After computing Q , we want to see if it exceeds the critical value of the corresponding chi-squared distribution. In this case, the p-value would be assessed for significance and if found below 0.05, this is evidence for residual correlation so we reject the null hypothesis. A large p-value over 0.05 indicates that the residuals do not exhibit autocorrelation, thereby satisfying the independence assumption (Shumway & Stoo, 2016). In this study we will be using the ACF plots to get a graphic picture of at the lags at which the autocorrelation is strongest while the results of Ljung-Box test will be used to confirm or deny the presence of autocorrelation.

Residuals Heteroscedasticity

The OLS assumes that all the residuals have constant variance across all values of the dependent variable. The distribution of the residuals should not show any systematic increase or decrease with the predicted values. When the variance is not constant (known as heteroscedasticity), OLS remains unbiased but inefficient and standard errors become inaccurate. This can cause problems with hypothesis tests and confidence intervals. In this study we use scale-location plots to check the homoscedasticity visually. The square root of the standardized residuals is plotted against the model's fitted values in these plots. A random scatter of points suggests that there is homogeneity of variance (homoscedasticity) whereas a systematic pattern, like a fan shape (increasing or decreasing spread) suggests heteroscedasticity.

To enhance the visual diagnosis, we will also conduct a formal statistical test to assess whether the variance of the residuals is dependent on the fitted values or other predictors. For that we will use the Breusch-Pagan test. The null hypothesis (H_0) of this test is that the error variance is constant and the alternative (H_1) assumes that it is heteroskedastic. The test regresses the squared residuals (e_i^2) on the fitted values from the original model:

$$e_i^2 = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon_i \quad (2.12)$$

The goal is to determine whether predictors contribute to a significant variation in the residual variance. After the regression, the Breusch-Pagan test statistic (LM) is calculated as:

$$LM = nR^2 \quad (2.13)$$

where:

- n is the number of observations
- R^2 is the coefficient of determination from the auxiliary regression

The statistic of the test follows a Chi-squared distribution with p degrees of freedom where p is the number of predictors used in auxiliary regression. A small p-value ($p < 0.05$) will result in rejecting the null hypothesis indicating heteroscedasticity and a large p-value indicates residuals are homoscedastic (Breusch & Pagan, 1979).

Normality of Residuals

Normality of residuals while not a strict requirement, it becomes essential as non-normal residuals may indicate model misspecification, outliers or other issues with the chosen variables, that can affect the reliability of the models inference. To assess the normality of residuals, we use a combination of visual tools - the QQ plot and the histogram. The quantile-quantile (QQ) plots, plot the residuals against the expected quantiles of a normal distribution and if residuals are normally distributed, the points will align closely with the diagonal line in the plot. Any systematic departures from the diagonal, such as an “S” shape, suggest skewness, while more dispersed points at the ends indicate heavy tails or kurtosis. A well-aligned plot suggests the residuals are approximately normal. In addition to the QQ plot a histogram of the residuals is plotted, looking for the desired bell-shaped curve, indicating that residuals are approximately normal.

Again, a formal statistical test will be used alongside the visual aids, to assess normality. The Shapiro-Wilk test in its null hypothesis (H_0) assumes that residuals follow a normal distribution while the alternative hypothesis (H_1) suggests that residuals are not normally distributed. The test statistic (W) is calculated as:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2.14)$$

where:

- $x_{(i)}$ is the i -th ordered sample value
- \bar{x} is the sample mean
- a_i are constants derived from the expected normal distribution

A large p-value ($p > 0.05$) indicates no significant evidence to reject H_0 , indicating normality and a small p-value ($p < 0.05$) suggests that the residuals deviate significantly from normality (Shapiro & Wilk, 1965).

Multicollinearity of Predictors

Multicollinearity refers to the case when two or more independent variables are highly correlated in a regression model. Due to multicollinearity, one predictor may be a linear combination of another predictor causing coefficient estimates to be unstable and making it difficult to interpret the effect of individual predictors. It can also lead to an increase in standard errors for the estimates of predictors which can decrease the coefficients' statistical significance. Even

though a model whose predictors display multicollinearity may produce accurate predictions, interpreting these predictions and test their results should be done carefully.

The diagnostic measure used to assess multicollinearity is called Variance Inflation Factor (VIF). VIF measures the boost in variance of one coefficient due to multicollinearity with other predictors. A regression is run for each independent variable X_i , whereby the variable X_i is the dependent variable and all other predictors are the independent variables. The R^2 value from this extra regression indicates how well other predictors manage to explain X_i . The VIF for X_i is calculated as:

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \quad (2.15)$$

where R_i^2 is the coefficient of determination from the auxiliary regression. According to the commonly accepted rule of thumb regarding VIF, values ≤ 5 indicate that there is low multicollinearity and values > 5 indicate moderate multicollinearity. The VIF results > 10 , are considered problematic, as they imply that this variable's variance is being inflated excessively due to other predictors. The Generalized VIF (GVIF) applies to a model that contains categorical predictors. In this case, the calculation allows for the degrees of freedom associated with the variable, so it can be used for categorical variables with multiple levels (Sheather, 2009).

2.4. Exponential Smoothing

Exponential smoothing models are a family of statistical models used for time series analysis and forecasting. These techniques assign progressively higher weights to recent observations, as the weights for the past observations exponentially decrease.

2.4.1. Simple Exponential Smoothing (SES)

Simple Exponential Smoothing (SES) works best for time series that do not exhibit any significant trend or seasonality over time. The method calculates forecasts based on previous observations by applying a weighted average where past observations get lesser weights while new observations get higher weights.

According to Hyndman et al. (2008), the SES model can be expressed as follows:

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \quad (2.16)$$

where:

- \hat{y}_{t+1} is the forecast for the next time period
- y_t is the actual value at time t
- \hat{y}_t is the forecast for time t
- α is the smoothing parameter ($0 < \alpha < 1$) that controls the weight

The model is simple yet effective for stable time series, reflecting the most recent changes in the forecast while still using previous data to influence predictions.

2.4.2. Holt-Winters Method

The Holt-Winters method (Triple Exponential Smoothing) extends SES and Holt's method³ to handle time series with both trend and seasonality. The model uses exponential smoothing but also separately considers level, trend and seasonality making it useful for datasets with recurring patterns over time. The level component of a given time series is calculated as:

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (2.17)$$

where l_t is the level at time t , y_t is the observed value, s_{t-m} is the seasonal component from m periods ago and α is the smoothing parameter for the level. Next, the equation for the trend component is:

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \quad (2.18)$$

where b_t is the trend at time t and β is the smoothing parameter for the trend.

³Holt's exponential method or Double Exponential Smoothing, extends SES by adding components to capture both level and trend, making it suitable for time series data with a consistent trend over time (Hyndman et al., 2008).

Moving forward, last component is the seasonal which is calculated as:

$$s_t = \gamma(y_t - l_t) + (1 - \gamma)s_{t-m} \quad (2.19)$$

where s_t is the seasonal component at time t and γ is the smoothing parameter for seasonality. Now the forecast equation can be computed by using the above in the equation:

$$\hat{y}_{t+h} = l_t + hb_t + s_{t+h-m(k)} \quad (2.20)$$

where h is the forecast horizon and $m(k)$ adjusts the seasonal component for periodicity. As seen by the calculations, by separately modeling level, trend and seasonality, this method creates predictions for short to medium term horizons. It is a fairly simple though effective model in handling complex time series structures (Hyndman et al., 2008).

2.5. Accuracy Evaluation

Following the modeling part, it is crucial to evaluate the accuracy of the model predictions in order to assess their performance and compare the results of different models. Different accuracy metrics are used as a mean to quantify the differences between predicted values and actual observations. In this study we will be using the scale-dependent error metrics MAE and RMSE and the percentage error metric MAPE.

Mean Absolute Error (MAE)

The MAE measures the average of predictions errors, providing a clear view of their magnitude and how far they are from the actual value. MAE is calculated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.21)$$

where n is the number of observations, y_i is the actual value at observation i and \hat{y}_i is the predicted value at observation i . It is one of the most common accuracy metrics and very simple (Neusser, 2016).

Root Mean Squared Error (RMSE)

RMSE is another scale-dependent error but in this case the calculation penalizes larger errors. This is achieved by squaring the differences thus making them sensitive to outliers:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.22)$$

where n is the number of observations, y_i is the actual value at observation i and \hat{y}_i is the predicted value at observation i (Neusser, 2016).

Mean Absolute Percentage Error (MAPE)

Last, we need a metric that allows us to not just judge the accuracy of a single model, but compare the results of different models. MAPE shows how accurate the overall predictions are as a percentage, making it easy to interpret and compare across different series and models. It is calculated as:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.23)$$

where n is the number of observations, y_i is the actual value at observation i and \hat{y}_i is the predicted value at observation i . We must note here that this metric can be sensitive to very small actual values, leading to inflated percentages (Shmueli & Lichtendahl, 2018).

3. Data Overview and Preprocessing

3.1. Overview of the Dataset

3.1.1. The M5 Competition

The M competitions are a series of open forecasting competitions, organized by the Makridakis Open Forecasting Center (MOFC) at the University of Nicosia, starting with the M1 contest in 1982. The main goal of the M competitions was to evaluate and compare the accuracy of different time series forecasting methods. The latest in the series was the M5 competition which took place in 2020 and invited submissions of forecasts from individuals and teams on retail sales data provided by the U.S. based mega-retailer, Walmart. M5 included two challenges: the *Accuracy Challenge*, which emphasized on point forecasts and the *Uncertainty Challenge*, which required estimating probabilistic sales distributions. These challenges reflected the practical needs of businesses to forecast not only expected sales but also the uncertainty associated with those predictions. The competition took place on Kaggle, a popular online platform for data science and machine learning, that attracted over 6,000 participants from around the world. Contestants were given a two-month timeline to submit their results and get an opportunity to win the cash prizes awarded to the top 5 submissions (Makridakis et al., 2022).

The dataset was provided by Walmart Inc., a multinational retail corporation and one of the world's largest companies by revenue, with operations in 18 countries and approximately 10,500 stores worldwide. Founded in 1962 and headquartered in Bentonville, Arkansas, Walmart operates a mix of discount department stores, grocery stores and e-commerce platforms. Renowned for its "Everyday Low Prices" strategy, Walmart serves millions of customers daily, offering a wide range of products including groceries, apparel, electronics and household goods (Walmart Inc., 2024). Walmart's efficient supply chain and focus on affordability have made them leaders in retail innovation and a key player in the global economy. The M5 competition focused on forecasting Walmart's daily sales across three U.S. states - California, Texas and Wisconsin and

10 stores over a 28-day period, from June 20, 2016, to July 17, 2016. There are two major hierarchies in the dataset, the *state-store* and the *category-department-item*. Participants relied on historical data spanning from January 2011 to June 2016, to train their models and create their predictions for each product, resulting in an impressive total of 3,049 time series.

3.1.2. Raw Data Overview

The provided dataset consisted of three primary files:

- **Sales Data:** Daily unit sales data at item level, hierarchically organized across states-stores and product categories-departments.
- **Calendar Data:** Information on each date, including weekdays, months and significant events like holidays (e.g., Super Bowl, Thanksgiving and Valentine's Day), which could influence consumer behavior.
- **Price Data:** Historical prices for all products, providing information on how pricing may affect demand.

These datasets combine internal factors, such as sales and prices with external influences, like holidays, allowing for a detailed analysis of the complex dynamics driving retail sales. Starting from the sales dataset, we have columns specifying the item code, department and category, alongside separate columns for the store and state hierarchies. The sales data itself is presented in a wide format, with a separate column for each date capturing daily sales figures. Moving to the pricing dataset, we find columns for the item, store, week and the corresponding price at which each item was sold. Finally, the calendar dataset provides a set of date related columns, including year, month, day, week and weekday, as well as indicators for U.S. holidays. Holidays are further categorized into Sporting, National, Cultural and Religious (event type) with an additional column naming the specific holiday (event name). Last we have three columns, indicating the days that the SNAP¹ benefits are loaded onto recipients cards for each particular state (CA, TX, WI).

As seen in Figure 3.1 sales are structured hierarchically, beginning at the highest level with the aggregated data that encompasses all elements. At the next level, the data is divided into three

¹The Supplemental Nutrition Assistance Program (SNAP), is a federal assistance program administered by the U.S. Department of Agriculture (USDA), that helps low-income individuals and families buy food. SNAP benefits are provided through a card, which can be used at authorized retailers like grocery stores and supports over 40 million Americans each year, while also boosting local economies through increased spending in food. More information regarding SNAP can be found here: <https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program>

states: California, Texas and Wisconsin. Each state contains multiple stores, with a total of ten stores represented across the dataset. Within each store, sales are categorized into three main product categories: FOODS, HOBBIES and HOUSEHOLD.

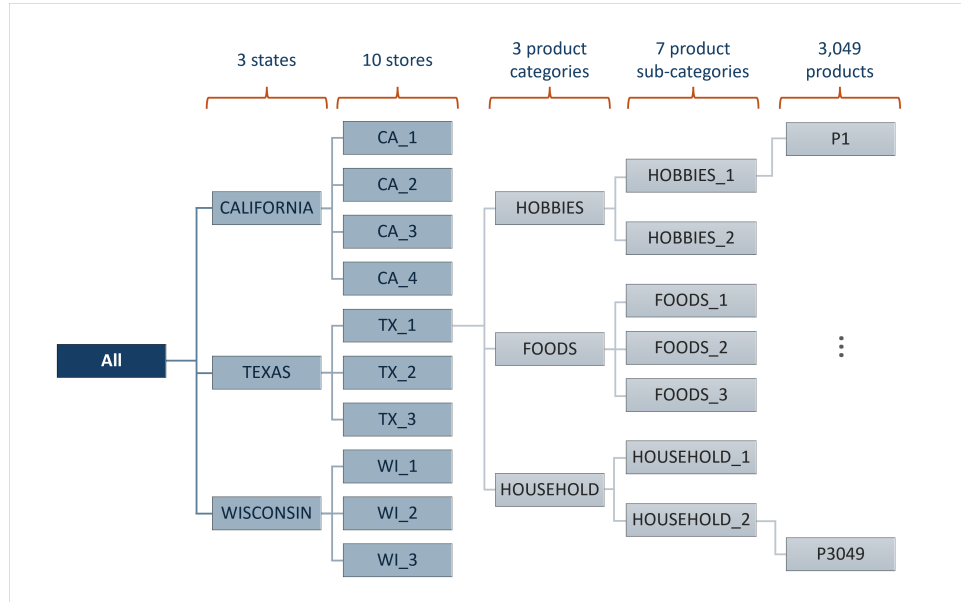


Figure 3.1.: Hierarchies of the M5 competition data

The group categories are further broken down into seven sub-categories (e.g., FOODS_1, FOODS_2 and FOODS_3) with the same structure for HOBBIES and HOUSEHOLD. The lowest level of the dataset contains 3,049 individual products. The structure of the data allows for analysis on various levels of sales, from broad aggregations to individual products at a store level. The features of the M5 dataset make it ideal for the objectives of this study which is to assess the demand patterns, to find out impactful variables affecting sales behavior and determine the effect of temporal patterns on retail sales. Furthermore, the dataset's applicability to actual retail operations makes it significant for achieving the broader goal of the dissertation, which is to demonstrate real-world practical examples for demand insights extraction.

3.2. Data Preprocessing

3.2.1. Combining the Data

Before we begin our analysis, we need to create a unified dataset that will include all available information from the given data files. The data engineering process must be done with extreme

attention to detail in order to maintain validity and avoid any errors that will alter the original information. Considering the above, we will begin our data preparation phase, in accordance to the forecasting framework described in Chapter 2.

Initially, we need to preprocess the datasets to prepare them for the join. Specifically, we will transform the calendar dataset by pivoting the SNAP-related columns into a long format. This transformation will create two new fields: one holding the state variable and one for the SNAP dummy (0/1). By restructuring the data, we create a unique date-state identifier in the calendar dataset, which will serve as keys for the subsequent dataset merges. Following the same process on the sales data and since each day of sales is held in a different column (wide format), we need to pivot them and replace them with two new fields, one for the day indication and another for the sales value. Since the sales and calendar datasets are now in the required format and the prices dataset does not need any modifications, we can proceed with merging all three datasets into a unified dataset joining them as depicted in Figure 3.2.

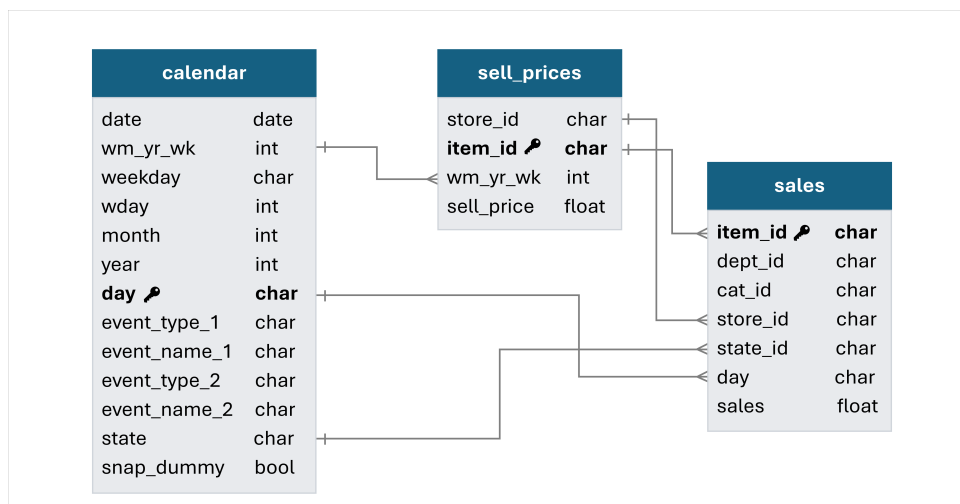


Figure 3.2.: Entity-relationship diagram of joins used for creating the unified dataset

The newly created dataset consolidates in a single structure all the categorical and numerical fields originally provided in the three separate files, streamlining our data wrangling process and providing all the necessary information for the next stage of preprocessing. In the following page, we present a table showcasing a random sample of 50 rows from this unified dataset.

item_id	dept_id	cat_id	store_id	state_id	day	sales	date	wm_yr_wk	weekday	wday	month	year	event_name_1	event_type_1	event_name_2	event_type_2	snap_dummy	sell_price
FOODS_3_763	FOODS_3	FOODS	TX_2	TX	d_1226	1	07/06/2014	11419	Saturday	1	6	2014	NA	NA	NA	NA	1	3.48
FOODS_2_012	FOODS_2	FOODS	TX_3	TX	d_474	2	16/05/2012	11216	Wednesday	5	5	2012	NA	NA	NA	NA	0	2.96
HOUSEHOLD_1_515	HOUSEHOLD_1	HOUSEHOLD	WI_2	WI	d_893	1	09/07/2013	11324	Tuesday	4	7	2013	Ramadan starts	Religious	NA	NA	1	2.47
FOODS_3_318	FOODS_3	FOODS	CA_3	CA	d_312	36	06/12/2011	11145	Tuesday	4	12	2011	NA	NA	NA	NA	1	1.28
FOODS_3_258	FOODS_3	FOODS	WI_2	WI	d_590	9	09/09/2012	11233	Sunday	2	9	2012	NA	NA	NA	NA	1	1.48
HOUSEHOLD_1_129	HOUSEHOLD_1	HOUSEHOLD	CA_1	CA	d_1663	2	18/08/2015	11529	Tuesday	4	8	2015	NA	NA	NA	NA	0	3.98
FOODS_2_055	FOODS_2	FOODS	TX_1	TX	d_1124	1	25/02/2014	11404	Tuesday	4	2	2014	NA	NA	NA	NA	0	2.00
HOBBIES_2_040	HOBBIES_2	HOBBIES	WI_1	WI	d_1683	1	07/09/2015	11532	Monday	3	9	2015	LaborDay	National	NA	NA	0	0.88
FOODS_1_024	FOODS_1	FOODS	CA_3	CA	d_153	2	30/06/2011	11122	Thursday	6	6	2011	NA	NA	NA	NA	0	3.97
HOUSEHOLD_1_496	HOUSEHOLD_1	HOUSEHOLD	CA_2	CA	d_1814	3	16/01/2016	11551	Saturday	1	1	2016	NA	NA	NA	NA	0	7.93
FOODS_3_215	FOODS_3	FOODS	TX_1	TX	d_1123	4	24/02/2014	11404	Monday	3	2	2014	NA	NA	NA	NA	0	1.24
HOBBIES_1_022	HOBBIES_1	HOBBIES	WI_3	WI	d_509	4	20/06/2012	11221	Wednesday	5	6	2012	NA	NA	NA	NA	0	7.18
FOODS_3_118	FOODS_3	FOODS	CA_1	CA	d_122	1	30/05/2011	11118	Monday	3	5	2011	MemorialDay	National	NA	NA	0	1.00
HOUSEHOLD_2_482	HOUSEHOLD_2	HOUSEHOLD	TX_1	TX	d_245	1	30/09/2011	11135	Friday	7	9	2011	NA	NA	NA	NA	0	5.94
FOODS_2_044	FOODS_2	FOODS	TX_2	TX	d_615	2	04/10/2012	11236	Thursday	6	10	2012	NA	NA	NA	NA	0	2.50
FOODS_2_233	FOODS_2	FOODS	CA_1	CA	d_1954	1	04/06/2016	11619	Saturday	1	6	2016	NA	NA	NA	NA	1	4.48
HOUSEHOLD_1_179	HOUSEHOLD_1	HOUSEHOLD	CA_3	CA	d_568	7	18/08/2012	11230	Saturday	1	8	2012	NA	NA	NA	NA	0	0.97
HOUSEHOLD_2_316	HOUSEHOLD_2	HOUSEHOLD	CA_3	CA	d_1004	1	28/10/2013	11340	Monday	3	10	2013	NA	NA	NA	NA	0	7.24
FOODS_3_039	FOODS_3	FOODS	TX_1	TX	d_1467	4	03/02/2015	11501	Tuesday	4	2	2015	NA	NA	NA	NA	1	3.98
FOODS_3_500	FOODS_3	FOODS	WI_1	WI	d_1968	3	18/06/2016	11621	Saturday	1	6	2016	NA	NA	NA	NA	0	2.50
FOODS_1_058	FOODS_1	FOODS	CA_1	CA	d_1284	1	04/08/2014	11427	Monday	3	8	2014	NA	NA	NA	NA	1	1.78
FOODS_3_568	FOODS_3	FOODS	CA_3	CA	d_1279	15	30/07/2014	11426	Wednesday	5	7	2014	NA	NA	NA	NA	0	1.00
FOODS_3_791	FOODS_3	FOODS	WI_1	WI	d_1876	1	18/03/2016	11607	Friday	7	3	2016	NA	NA	NA	NA	0	2.50
FOODS_3_029	FOODS_3	FOODS	CA_3	CA	d_369	2	01/02/2012	11201	Wednesday	5	2	2012	NA	NA	NA	NA	1	1.28
HOUSEHOLD_1_194	HOUSEHOLD_1	HOUSEHOLD	CA_2	CA	d_1826	1	28/01/2016	11552	Thursday	6	1	2016	NA	NA	NA	NA	0	4.98
HOUSEHOLD_1_094	HOUSEHOLD_1	HOUSEHOLD	WI_2	WI	d_1825	1	27/01/2016	11552	Wednesday	5	1	2016	NA	NA	NA	NA	0	3.48
HOBBIES_1_339	HOBBIES_1	HOBBIES	CA_1	CA	d_1892	2	03/04/2016	11610	Sunday	2	4	2016	NA	NA	NA	NA	1	7.98
HOUSEHOLD_1_179	HOUSEHOLD_1	HOUSEHOLD	CA_3	CA	d_478	14	20/05/2012	11217	Sunday	2	5	2012	NA	NA	NA	NA	0	0.97
FOODS_2_285	FOODS_2	FOODS	WI_3	WI	d_1112	1	13/02/2014	11402	Thursday	6	2	2014	NA	NA	NA	NA	0	3.98
FOODS_3_243	FOODS_3	FOODS	CA_1	CA	d_19	1	16/02/2011	11103	Wednesday	5	2	2011	NA	NA	NA	NA	0	1.48
FOODS_3_290	FOODS_3	FOODS	CA_2	CA	d_1963	2	13/06/2016	11620	Monday	3	6	2016	NA	NA	NA	NA	0	1.00
FOODS_3_591	FOODS_3	FOODS	CA_4	CA	d_585	3	04/09/2012	11232	Tuesday	4	9	2012	NA	NA	NA	NA	1	2.48
HOUSEHOLD_1_389	HOUSEHOLD_1	HOUSEHOLD	TX_2	TX	d_1126	6	27/02/2014	11404	Thursday	6	2	2014	NA	NA	NA	NA	0	0.94
FOODS_3_046	FOODS_3	FOODS	TX_3	TX	d_682	1	10/12/2012	11246	Monday	3	12	2012	NA	NA	NA	NA	0	1.98
HOBBIES_1_203	HOBBIES_1	HOBBIES	TX_2	TX	d_1371	1	30/10/2014	11439	Thursday	6	10	2014	NA	NA	NA	NA	0	12.98
FOODS_2_371	FOODS_2	FOODS	CA_3	CA	d_392	6	24/02/2012	11204	Friday	7	2	2012	NA	NA	NA	NA	0	1.97
FOODS_1_177	FOODS_1	FOODS	TX_1	TX	d_1014	2	07/11/2013	11341	Thursday	6	11	2013	NA	NA	NA	NA	1	8.48
HOUSEHOLD_2_133	HOUSEHOLD_2	HOUSEHOLD	CA_3	CA	d_1339	3	28/09/2014	11435	Sunday	2	9	2014	NA	NA	NA	NA	0	4.97
HOBBIES_1_030	HOBBIES_1	HOBBIES	CA_4	CA	d_1029	4	22/11/2013	11343	Friday	7	11	2013	NA	NA	NA	NA	0	0.70
HOUSEHOLD_1_218	HOUSEHOLD_1	HOUSEHOLD	WI_1	WI	d_1928	1	09/05/2016	11615	Monday	3	5	2016	NA	NA	NA	NA	1	1.94
FOODS_3_304	FOODS_3	FOODS	TX_2	TX	d_123	4	31/05/2011	11118	Tuesday	4	5	2011	NBA Finals Start	Sporting	NA	NA	0	3.48
FOODS_3_531	FOODS_3	FOODS	WI_1	WI	d_1619	4	05/07/2015	11523	Sunday	2	7	2015	NA	NA	NA	NA	1	2.48
HOUSEHOLD_1_426	HOUSEHOLD_1	HOUSEHOLD	WI_2	WI	d_1655	1	10/08/2015	11528	Monday	3	8	2015	NA	NA	NA	NA	0	2.58
FOODS_2_299	FOODS_2	FOODS	CA_4	CA	d_492	1	03/06/2012	11219	Sunday	2	6	2012	NA	NA	NA	NA	1	2.00
HOUSEHOLD_2_184	HOUSEHOLD_2	HOUSEHOLD	CA_4	CA	d_829	2	06/05/2013	11315	Monday	3	5	2013	NA	NA	NA	NA	1	1.96
HOBBIES_1_122	HOBBIES_1	HOBBIES	TX_2	TX	d_918	1	03/08/2013	11328	Saturday	1	8	2013	NA	NA	NA	NA	1	1.63
HOUSEHOLD_1_061	HOUSEHOLD_1	HOUSEHOLD	CA_4	CA	d_1425	2	23/12/2014	11447	Tuesday	4	12	2014	NA	NA	NA	NA	0	1.88
FOODS_3_218	FOODS_3	FOODS	WI_2	WI	d_1111	2	12/02/2014	11402	Wednesday	5	2	2014	NA	NA	NA	NA	1	2.00
FOODS_3_065	FOODS_3	FOODS	WI_3	WI	d_1306	1	26/08/2014	11430	Tuesday	4	8	2014	NA	NA	NA	NA	0	2.98
FOODS_2_149	FOODS_2	FOODS	WI_3	WI	d_1413	5	11/12/2014	11445	Thursday	6	12	2014	NA	NA	NA	NA	1	6.98

Table 3.1.: Sample rows from the joined dataset

3.2.2. Data Aggregation

The M5 Forecasting Competitions' goal was to generate daily sales forecasts at product level. While this level of granularity serves the objectives of the competition, it does not align with the scope of our research questions or the broader strategic aims of this study, which focus on uncovering larger demand patterns and their implications for planning and forecasting.

To address the research questions outlined in Chapter 1, this study aggregates daily sales data to a weekly frequency and narrows the analysis to the three product categories: FOODS, HOBBIES and HOUSEHOLD. Weekly aggregation reduces the noise inherent in daily data while retaining important patterns, such as trends, seasonality and spikes caused by events, which are essential for effective planning. Daily data, are often too volatile and influenced by short-term fluctuations, making it less practical for uncovering meaningful long-term insights. On the other hand, moving to a monthly frequency would obscure crucial details, such as weekly cycles and promotional impacts, while drastically reducing the number of observations available for analysis. Weekly aggregation strikes the right balance between smoothing daily noise and maintaining sufficient granularity for actionable insights and this frequency will be used consistently throughout the project.

Focusing on product categories rather than individual items simplifies the complexity of working with more than 3,000 products and reduces computing needs for calculations and modeling. Analyzing data at the category level provides a broader perspective on consumer behavior and supply chain dynamics. Working in a category level we can capture meaningful insights for demand patterns, making this level of aggregation both practical and strategically important. This approach also ensures alignment with the real-life tasks of a demand planner's role in a big organization like Walmart. Weekly, category-level forecasts offer significant insights that can support both strategic and tactical planning while at the same time ensures that the study addresses the key research questions.

3.2.3. Feature Engineering

Following the decision to use weekly frequency by product category, some data engineering work is needed to adapt the daily data features to the weekly aggregated level. To begin with, we aggregate the sales by category from Monday to Sunday daily. This means we sum the daily sales to derive total sales by category with weekly frequency. Therefore, we create a dataset with three time series of weekly data with 282 observations each. Next, we calculate each week average sell price to use as a numerical field in the dataset.

To include SNAP information, we generate a new variable that counts the number of SNAP days in each week. We can similarly group the columns of the events to create a new field that counts the number of events in the week. While these transformations simplify the original structure, they retain useful information from the event and SNAP data, which can be later used as explanatory variables.

At this first preprocessing stage the weekly dataset includes the following columns:

- Week
- Category
- Sales
- Mean Price
- Events Count
- SNAP Days Count

These features provide a solid foundation for the subsequent data exploration sections and gives us the ability to further develop new features as needed.

4. Exploratory Data Analysis (EDA)

4.1. Descriptive Statistics

The dataset prepared in Chapter 3 contains a mix of numerical and categorical variable types. Starting our EDA part of the project, we begin by calculating basic statistics for the two primary numerical variables in the dataset - sales and price - results of which are presented in Table 4.1.

	FOODS	HOBBIES	HOUSEHOLD
	(N=282)	(N=282)	(N=282)
Sales			
Mean (SD)	165,778.50 (24,531.66)	22,538.91 (3,772.39)	53,371.65 (12,291.09)
Median [Min, Max]	170,226.50 [45,936, 225,812]	22,630.50 [7,121, 30,690]	55,610.50 [11,323, 78,291]
IQR [25%, 75%]	30,375.00 [152,274.50, 182,649.50]	5,738.50 [19,752.00, 25,490.50]	18,750.00 [43,354.00, 62,104.00]
Skewness	-0.73	-0.22	-0.43
Kurtosis	1.41	-0.12	-0.59
CV	0.15	0.17	0.23
Missing	0	0	0
Price			
Mean (SD)	2.98 (0.10)	5.16 (0.43)	4.76 (0.09)
Median [Min, Max]	2.96 [2.77, 3.21]	5.30 [4.16, 5.78]	4.75 [4.54, 5.06]
IQR [25%, 75%]	0.14 [2.91, 3.05]	0.63 [4.86, 5.49]	0.11 [4.70, 4.81]
Skewness	0.38	-0.78	0.78
Kurtosis	-0.67	-0.70	0.41
CV	0.03	0.08	0.02
Missing	0	0	0

Table 4.1.: Summary statistics for the continuous variables

Our aggregated data consist of 282 observations for each category with no missing values. With the largest mean sales (166,778), FOODS is the dominant category in terms of sales volume, followed by HOUSEHOLD (53,372) and HOBBIES (22,539). For all categories mean and median sales values are close, suggesting a fairly symmetric distribution. The relatively high standard deviation (SD) for FOODS and HOUSEHOLD combined with their interquartile range (IQR) results, indicate a moderate variability in sales compared to the more consistent sales of HOBBIES where the variation appears to be minimal and sales patterns appear to be more consistent. Regarding the shape of the sales probability distribution FOODS has a slightly left-tailed skew with

a near-normal kurtosis and a low Coefficient of Variation (CV), indicating a relatively symmetric and consistent sales pattern with moderate variability. HOBBIES has the lowest skewness and kurtosis, with a very symmetric and flat distribution. The CV is slightly higher, indicating more relative variability despite the absolute SD being the smallest. HOUSEHOLD has relatively higher variability in relation to its mean (higher CV) with moderate kurtosis indicating near-normality.

Regarding the Price statistics, for FOODS we have almost identical mean and median pointing to a symmetric distribution with a narrow range and a small SD indicating a consistent pricing pattern. The slight positive skewness tells us that there might be some weeks with higher mean price and kurtosis suggests a distribution without extreme outliers. CV is the lowest showing that FOODS mean weekly prices are stable. Moving on to HOBBIES, we have the widest range among the categories, with a high SD and IQR, indicating diverse pricing probably due to a mix of weeks with very low and higher mean prices. The same can be deducted from longer tail on the lower end in skewness and a CV higher than the other two categories. Finally, HOUSEHOLD category demonstrates similar pricing patterns with FOODS, with symmetrical distribution and despite the small positive skewness it has the highest stability in pricing ($CV = 0.02$).

4.2. Distributional Analysis

Following the insights gathered from descriptive statistics we will visualize the distributions of both weekly sales and weekly mean price, in order to better understand the shape, spread and variability by category.

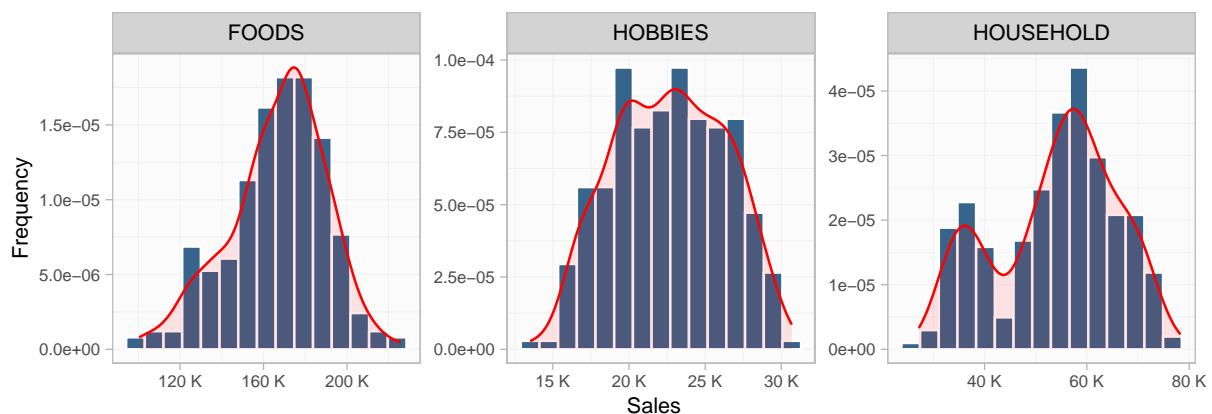


Figure 4.1.: Distribution plots of the sales variable

As seen in Figure 4.1, the majority of the FOODS sales are concentrated between 150K and 200K, with fewer occurrences below 120K or above 200K. The slight skew suggests occasional weeks with lower-than-average sales but overall we have a symmetrical shape. For HOBBIES, the flatter density curve and wider spread indicate a more evenly distributed range of sales compared to FOODS, though the distribution is less variable. Now for HOUSEHOLD, the two-peak shape, suggests the possibility of two distinct sales patterns, which could be driven by factors like seasonal effects or varying demand within the category.

From the FOODS category violin plot in Figure 4.2 we see a symmetric and concentrated shape indicative of stable sales, although we have a small number of outliers visible at the lower tail (below 120K), suggesting weeks with significantly reduced sales. The violin plot for HOBBIES is also symmetric, but the distribution is flatter and less concentrated, reflecting more consistent sales across its range (15K–30K) and the upper and lower tails are less pronounced compared to the other two categories. Only a few points of outliers suggest that HOBBIES sales are highly consistent week-to-week, with no extreme deviations. Again, the bi-modal pattern is visible for HOUSEHOLD category, with a few non significant points outside the lower tail. This shape might indicate different demand cycles for the category.

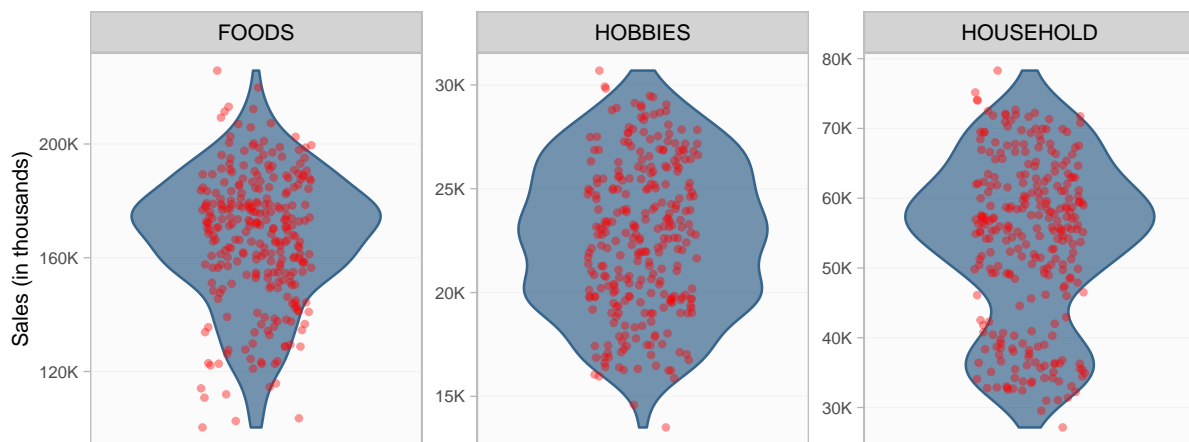


Figure 4.2.: Violin plots of weekly sales by category

Moving on to mean weekly price histograms in Figure 4.3, we have a tight and symmetric distribution for FOODS, with the overlaid density curve following closely the bars, indicating no significant outliers. The slight positive skewness observed may reflect occasional weeks with marginally higher prices. For HOBBIES the histogram reveals a bi-modal shape, with a broader distribution compared to FOODS, indicating diversity in the mean price of some weeks in this category. Lastly, similar to FOODS, HOUSEHOLD demonstrates symmetric and con-

sistent pricing, with a slightly positive skewness due to occasional higher mean prices for some weeks.

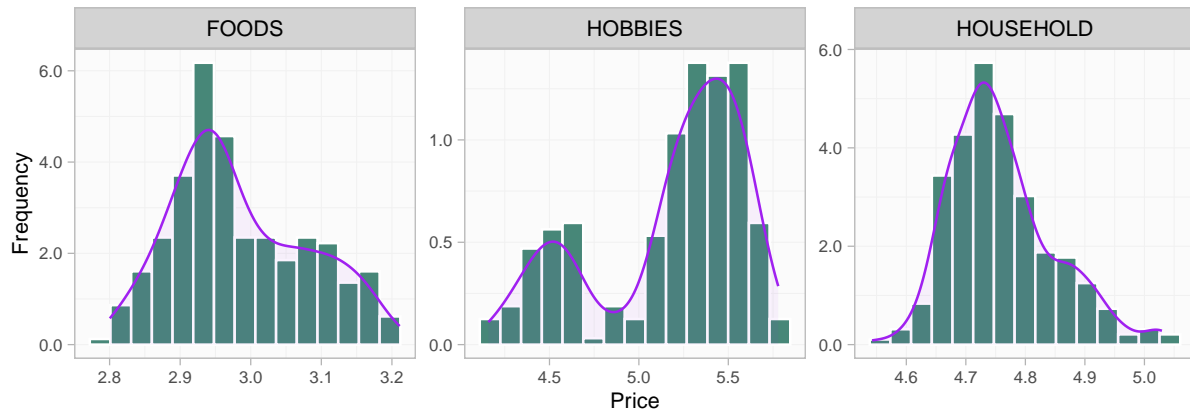


Figure 4.3.: Distribution plots of the price variable

Consistent with the above, from the price violin plots in Figure 4.4, we have a moderate spread for FOODS with no extreme widening or narrowing, suggesting that prices are consistent. HOBBIES confirms the two-peak price distribution, with observations concentrated in specific lower and higher price ranges. Last, HOUSEHOLD prices shape with more points at the lower part and fewer in the middle, reflect the skewness we see in the histogram.

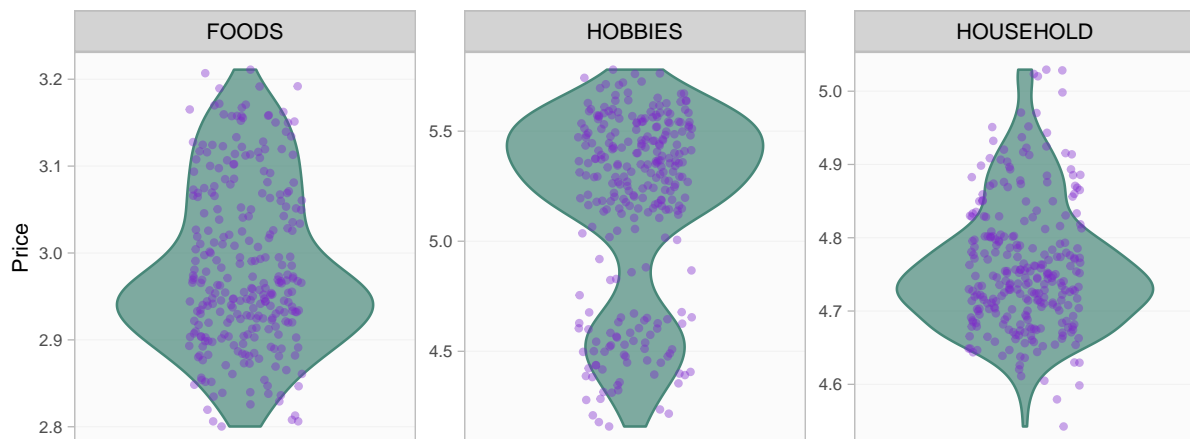


Figure 4.4.: Violin plots of weekly mean price by category

Summing up the distribution analysis for the two continuous variables in the weekly dataset, we can overall derive that FOODS category appears to be the most stable both in sales and pricing, while HOBBIES and HOUSEHOLD exhibit more variability in their results. This may suggest some kind of dynamic behavior for these two categories that we will try to explore further.

4.3. Categorical Exploration

For this next part we will dive into the categorical variables, visualize their totals and explore their influence on sales. Confirming the conclusions from previous parts, FOODS with over 46 million units sold, is the category that holds the largest volume of sales across the time span,

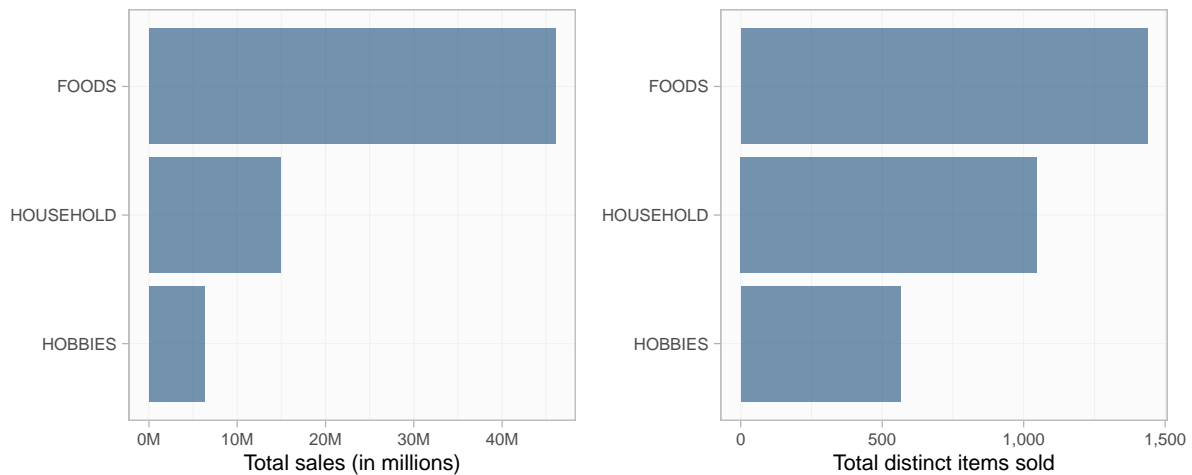


Figure 4.5.: Total units sold and total distinct products sold by category

followed from distance by HOUSEHOLD category (15 million), with HOBBIES category having the lowest sales volume (6 million). These results align with the expected commercial role of a grocery store chain, where customers primarily shop for essential food and household products on a weekly basis. This also can explain why a category like hobbies is less significant, as it likely serves as a complementary category to the other two.

Regarding the SNAP benefits variable, according to calculations on the original data there are 15 days each month when the beneficiaries receive the food stamp credits in their cards. This staggered schedule is based on the last digits of individuals' social security numbers, designed to distribute demand evenly within the month. Exploring the impact of SNAP card loading days in sales we get some interesting insights from Figure 4.6, where there is a clear trend in FOODS sales as the number of SNAP days increases. In addition it seems that the range of sales is relatively narrow for fewer SNAP days (0-2) but becomes wider as SNAP days increase, suggesting higher variability in spending during weeks with more SNAP days. Unlike FOODS, the other two categories show no obvious patterns that would suggest significant impact of SNAP days on their sales, likely because the items in these categories are less essential and probably are not included in the program.

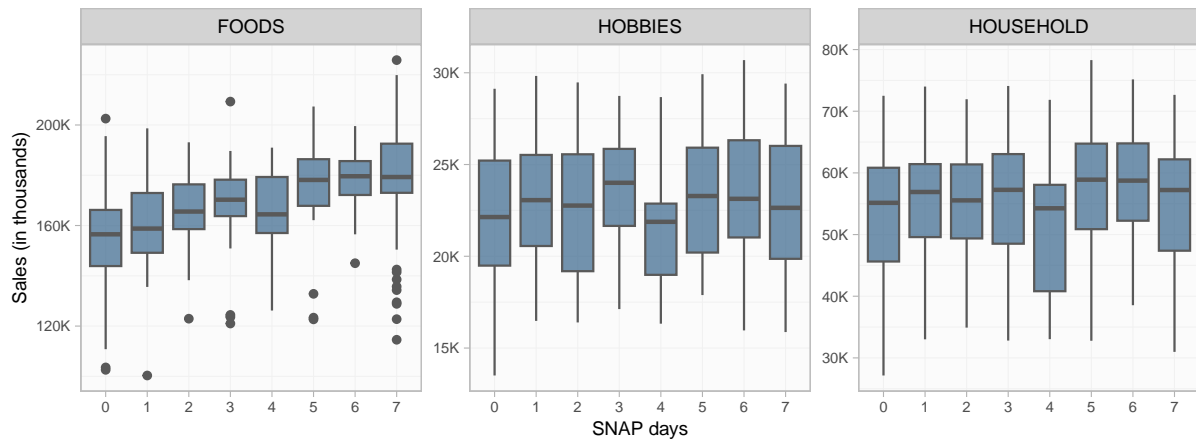


Figure 4.6.: Boxplots of weekly sales by number of SNAP days in a week

Advancing our EDA to Events, the other categorical variable in our dataset, we calculate a total of 31 distinct holidays marked in the data, 11 of them being “National”, 10 occasions are “Religious”, 7 are “Cultural” and 3 are “Sporting” events. Again using boxplots to visually assess the impact of each event category on sales, we see that for HOBBIES and HOUSEHOLD sales show minimal variation across event types suggesting that these categories remain relatively unaffected by events. On the other side, although the variation is small, FOODS appears to be more sensitive especially during sporting events where we see a rise and during national holidays where we have a slight negative impact.

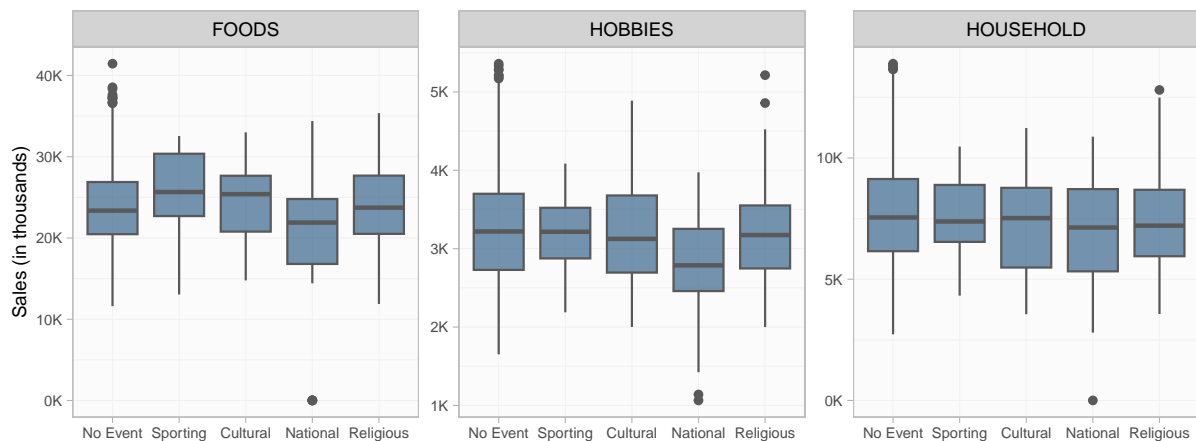


Figure 4.7.: Boxplots of weekly sales by event type

Overall we observe that both Events and SNAP variables have some impact on sales especially on the FOODS category which seems to be more sensitive. The above could be found useful for our modeling process later on.

4.4. Cross-Sectional Analysis

For this part of the EDA we will further explore relationships between the variables of the weekly dataset, trying to uncover significant interactions that could provide valuable insights for understanding the data. We will utilize visual representations of these interactions in an attempt uncover details that can be interpretable and help our modelling process later.

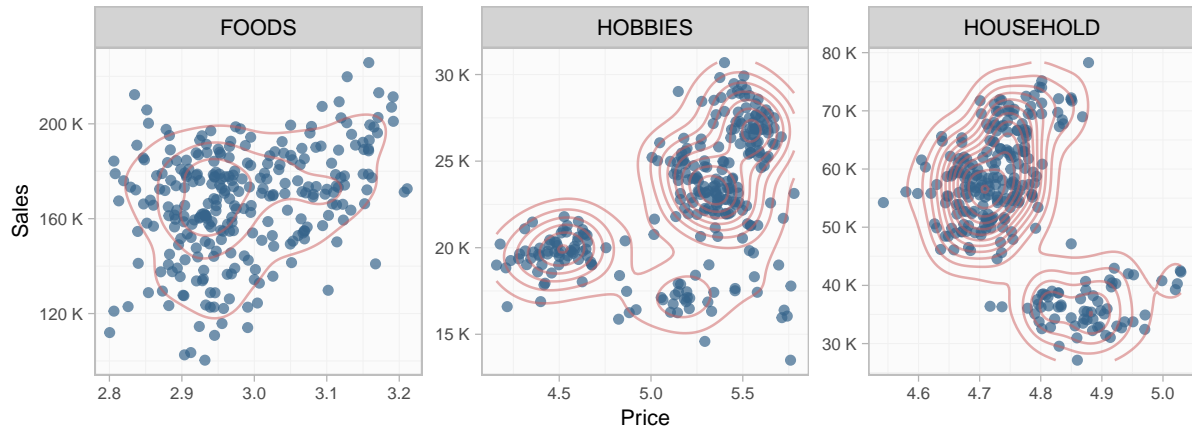


Figure 4.8.: Scatter plots of weekly sales vs weekly mean price

The contour lines in the FOODS scatter plot in Figure 4.8, produced as a result of a two-dimensional kernel density estimation, suggest a dense concentration of data points with a weak positive relationship between Price and Sales. Price increases have a slight positive effect on sales, but the relationship is not strong indicating that FOODS sales are likely driven by consistent demand. In HOBBIES, we see distinct clusters of sales at different price ranges with a stronger positive relationship between price and sales compared to FOODS. The above suggest that weeks with higher mean price tend to go along with higher sales. This aligns with the diversity in price distribution we have seen previously. Last, in HOUSEHOLD we have two main clusters showing a weak negative relationship with weeks of higher price and reduced sales. This pattern may reflect seasonal or changes in the assortment of the category.

As a next step we will quantify the relationship of all continuous variables by computing their correlation coefficients and visualize them into correlogram plots in Figure 4.9. The results for the FOODS category indicate a positive correlation of sales with SNAP days (0.45) and a less strong with price (0.33). For HOBBIES, sales seem to be most influenced by price (0.52) and for HOUSEHOLD sales correlate negatively with price (-0.47), indicating that higher prices tend to reduce sales while other variables show non-significant correlations.

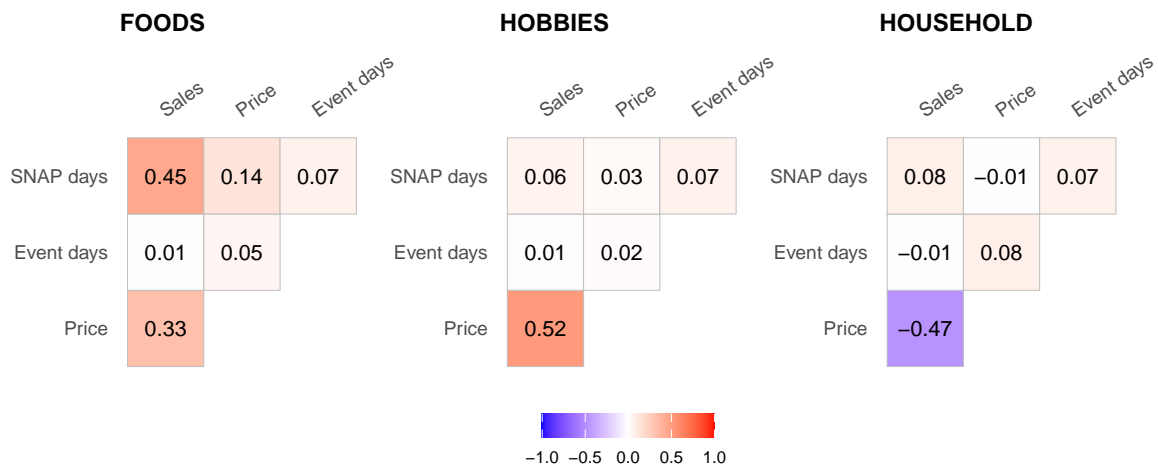


Figure 4.9.: Correlation of numerical variables by category

4.5. Key Takeaways

In this chapter we established an initial understanding of the weekly aggregated dataset and the basic attributes for each variable. Looking closely at sales and its connection with other variables we have laid the groundwork for recognizing trends that affect the behavior of all three categories. From the early descriptive statistics, FOODS was identified as the leading category in sales volume, followed by HOUSEHOLD and HOBBIES. The distributions of data showed patterns to be pretty symmetric with some few outliers. Next, it seems that weeks with more than two SNAP days show some increase in sales, especially in FOODS, indicating an important factor. However, Events appear to be relatively weakly correlated with sales, as we will need to investigate further in the next parts.

These early findings call for different strategies for each category, depending on their unique features. In addition, the associations and distributions seen previously can aid in further analysis starting from the next chapter where we will focus on the time series characteristics of our data.

5. Time Series Analysis

5.1. Time Series Overview

The weekly time series of the three product categories FOODS, HOBBIES and HOUSEHOLD are depicted in Figure 5.1, showcasing the distinct behaviors over the given period of approximately five years, from week 4 in January 2011, to week 24 in June 2016.

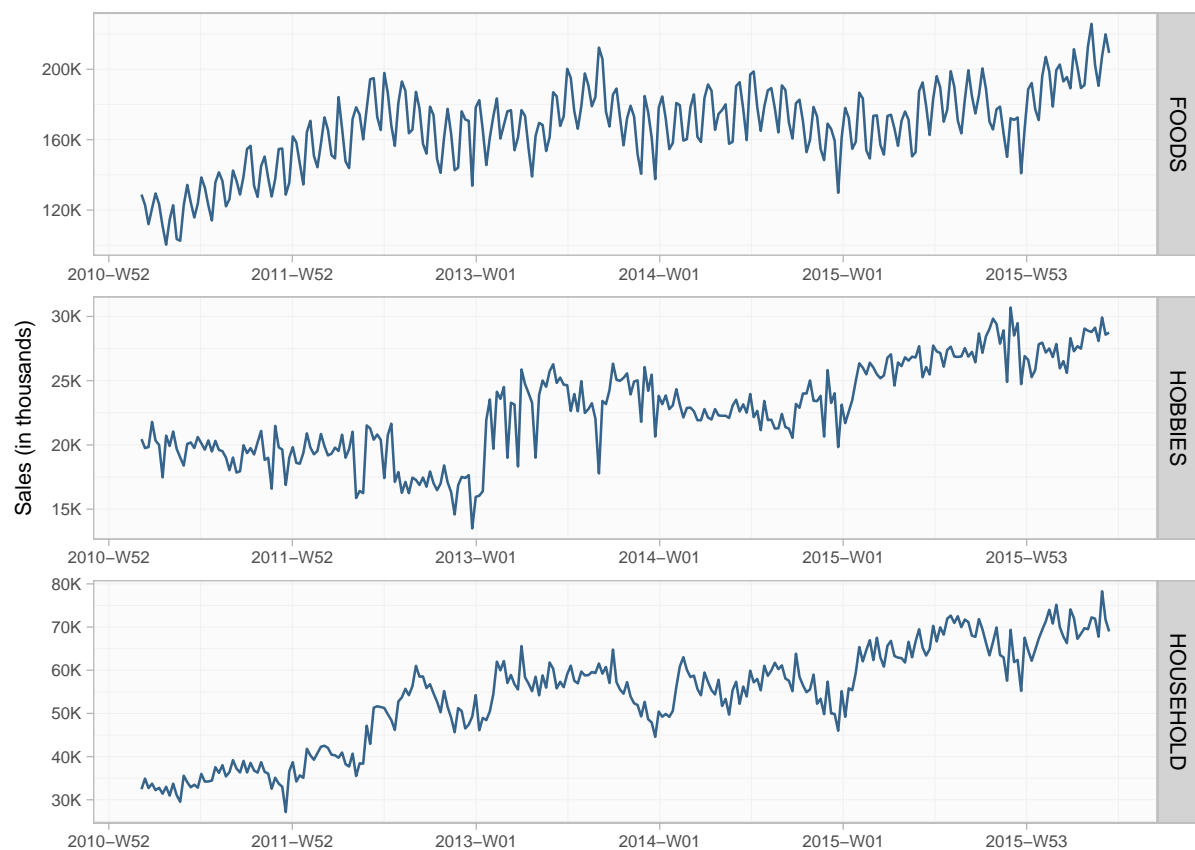


Figure 5.1.: Time series plots of weekly sales over time

5.1.1. FOODS

The time series for FOODS shows a distinct upward trend, indicating growing demand over time. A particularly strong upward movement is visible throughout 2011, suggesting a phase of expansion, which begins to stabilize in 2012. During this we observe the emergence of a clear annual seasonality that is consistent over the remainder of the period. At the start of 2016, a new trend emerges, driving the sales line upwards. Focusing on cyclical behavior, starting from 2012, a consistent pattern of peaks and troughs can be seen each year, suggesting that consumer behavior for this category is closely tied to annual cycles, likely influenced by recurring events like holidays and seasonal demand. The variation between the seasonal highs and lows remains relatively stable over the years, indicating that the impact of seasonality on FOODS sales does not diminish despite the overall upward trend. In addition to the annual seasonality, there also appears to be evidence of shorter-term cycles within the year, possibly linked to monthly fluctuations in demand.

5.1.2. HOBBIES

The sales of HOBBIES present a trending behavior as well, although more irregular compared to FOODS. Changes in levels are evident, with noticeable periods of plateau or sharp transitions, reflecting instability or potential external influences on demand. For example, in 2012 and during the whole year the level is completely changed after some shocks that are visible in the end of 2011. Then again after some intense spikes and dips at the end of 2012 the level is again changing, leading to a period of high volatility for the most part of 2013. These sharp declines and peaks suggest significant disruptions or sudden drops in demand, possibly tied to supply chain interruptions, or other external shocks. Despite the volatility, some indications of potential seasonal behavior exist, although they are less consistent than in FOODS. Certain peaks at similar times each year, hint at seasonal demand cycles, however the amplitude and timing of these seasonal peaks vary significantly, indicating that any cyclic behavior is influenced by additional factors. Overall, the combination of frequent level shifts, high volatility and inconsistent seasonality suggests that the HOBBIES time series is a challenging case and needs further understanding.

5.1.3. HOUSEHOLD

The HOUSEHOLD series exhibits a gradual upward trend over the observed period indicating a consistent slower increase in demand over the years. While the upward movement is evident, it does not demonstrate the sharp accelerations seen in the early years in FOODS, suggesting that growth in this category may be more stable. From the beginning in 2011 up to the end of 2012, there appears to be a phase when sales are not yet stabilized, growing in an irregular manner. After 2013, patterns start being more consistent and trending behavior becomes more stable, reflecting a maturing or more predictable demand over the later years. Seasonality in this series is apparent, but it lacks the prominence and regularity observed in FOODS. Peaks and troughs appear periodically, indicating cyclical behavior, though with a smaller amplitude. This weaker seasonality suggests that demand for household products is less tied to events or holidays and may be characterized more by steady consumption patterns. The overall variance in the series remains relatively consistent, with no apparent structural breaks or abrupt shifts in the underlying sales behavior. This stability in variance suggests that the HOUSEHOLD category does not experience the same volatility or sudden level changes seen in HOBBIES, making it more predictable.

5.2. Seasonality Visualization

For the next part we will deep dive into the cyclical component exploration in our time series, searching for recurring patterns that will confirm the presence of seasonality. This is a crucial part in our data exploration process since, we have already uncovered signs of seasonal patterns in Chapter 4, that could be confirmed in this section.

Starting from Figure 5.2, we have a visual of the monthly patterns by year. Across all three product categories, the annual cycles show moderate intensity, with relatively consistent variability between months. However, the year-on-year lines reveal the presence of annual cycles, with noticeable peaks in specific months such as October and December. For FOODS, aside from the year 2011, the patterns and levels remain closely aligned across all other years, indicating strong consistency. In contrast, HOBBIES and HOUSEHOLD display more dispersed patterns, reflecting greater variability in levels and stronger shifts in trends over time, as we have already highlighted earlier. Interestingly, we see that FOODS and HOBBIES share similar annual behavior on the other side, HOUSEHOLD seems to differ, with peaks during summer months instead of the months of the last quarter.

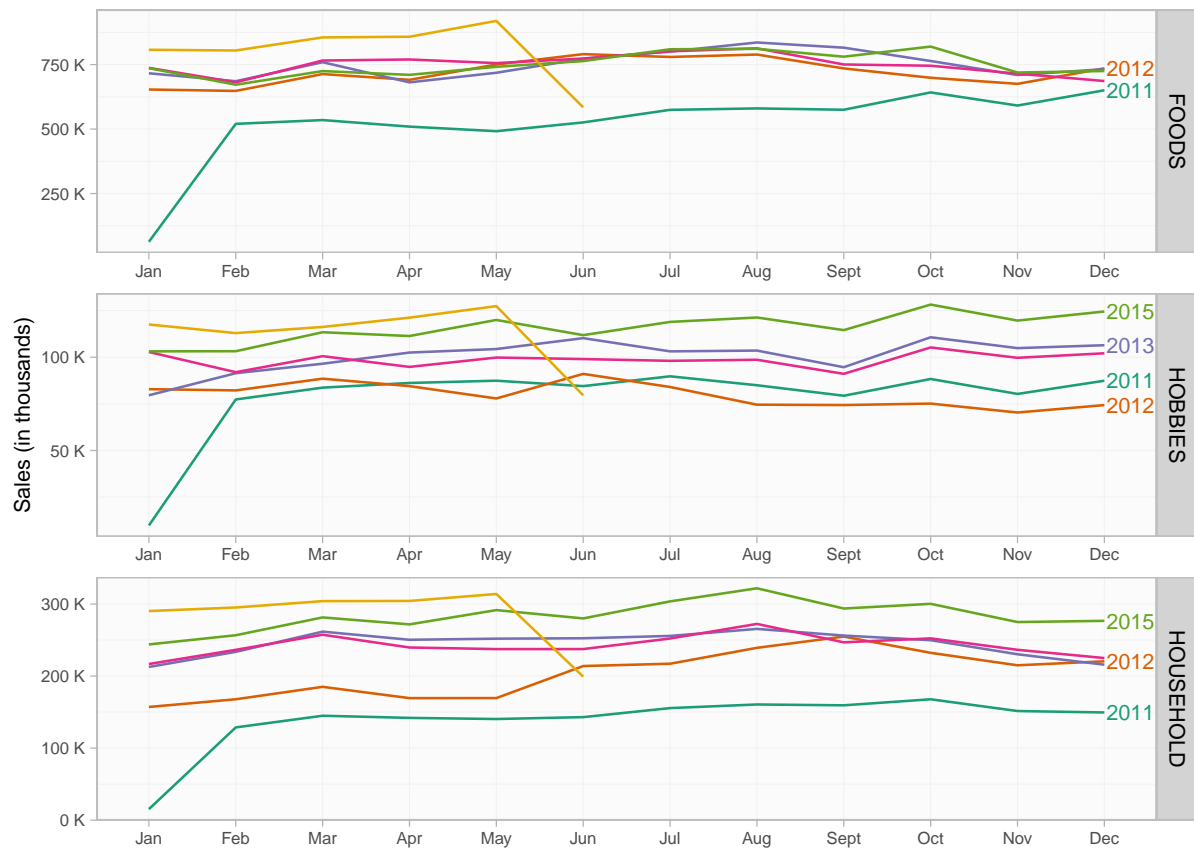


Figure 5.2.: Seasonal plots of monthly sales

Moving on to Figure 5.3, we have the sales patterns by week-of-year this time. From this breakdown of our time series, we see a more detailed picture of sales over time but similar to the monthly plot. For the FOODS category, with the exception of the very low level for 2011 and the higher for 2016, we see once more, the same stable and consistent patterns. HOBBIES shows irregular seasonality with patterns being more similar towards the end of each year, while being more volatile compared to FOODS. HOUSEHOLD on the other side, shows a moderate seasonal behavior with some consistency but weaker amplitude than FOODS.

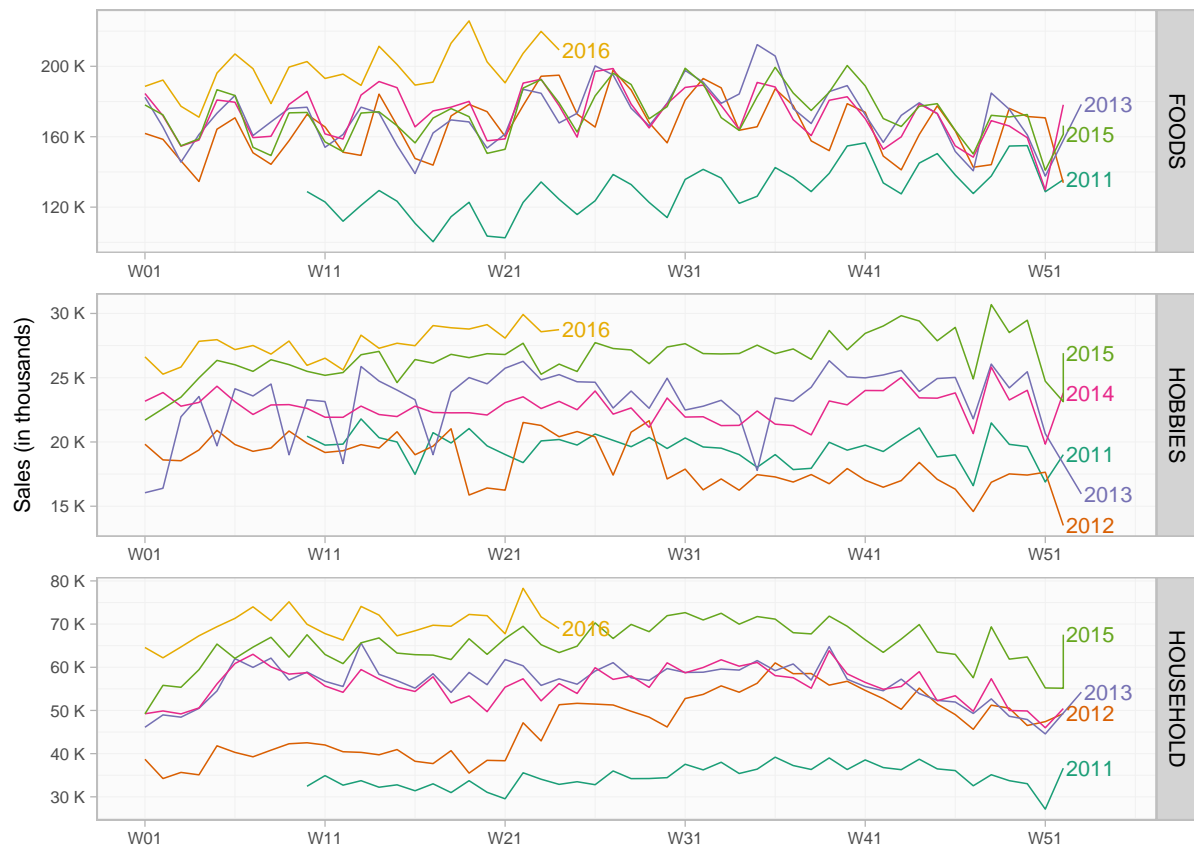


Figure 5.3.: Seasonal plots of weekly sales

Overall, from the first two visuals it is evident that the series exhibit an annual cycle and most likely it is also suggested the presence of a secondary week-of-month seasonality, especially in the FOODS category.

To further explore this secondary cycle, we will create boxplots of sales grouped by the week-of-month for each product category. In Figure 5.4 we see a noticeable variability in FOODS sales across each week of month. Sales are gradually declining as weeks progress towards the end of the month. This suggests that there is indeed a secondary week-of-month pattern, likely tied to consumer behavior such as paycheck cycles or early-month replenishment. In contrast to FOODS, the sales for HOBBIES remain relatively stable across all weeks of the month.

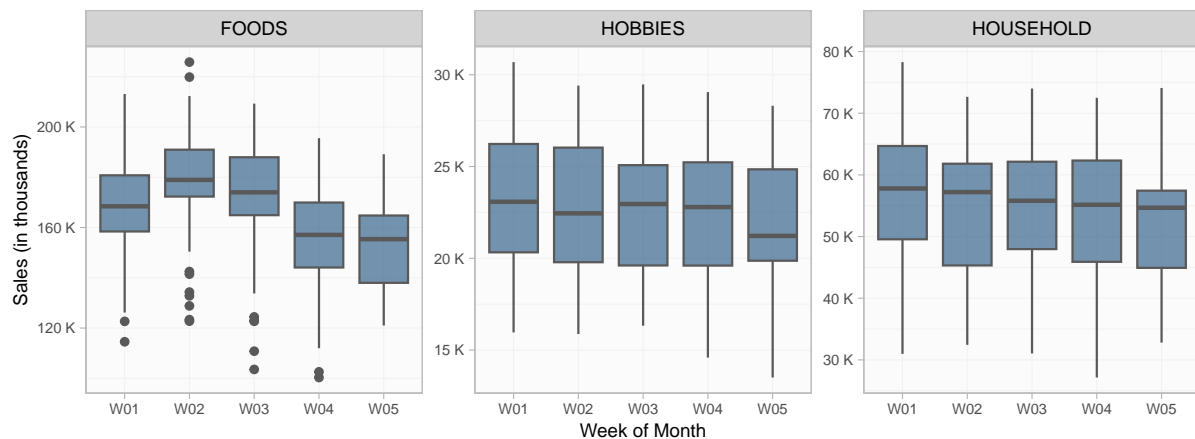


Figure 5.4.: Boxplots of sales by week-of-month

The medians and interquartile ranges are consistent, suggesting that this category does not exhibit a strong week-of-month effect. For HOUSEHOLD, sales are somewhat higher in W01, followed by a steady decline in the subsequent weeks, though the difference is not as pronounced as in FOODS. This could indicate a moderate week-of-month seasonality, where consumers prioritize household purchases earlier in the month, possibly as part of planned monthly budgets and similar to what we have seen already with FOODS.

5.3. Time Series Decomposition

Building on the seasonal exploration of previous sections, we will apply a time series decomposition method to separate the series into their components: trend, seasonality and residuals. By identifying and separating these components, we can assess how much of the sales variation is driven by seasonality and how consistent the trend is across the different product categories.

The FOODS decomposition model's results plot seen in Figure 5.5 demonstrates a long-term trend component which appears to be steadily increasing during the entire observed time period. It starts at about 120K and rises to about 190K at the end.

The third panel shows a strong seasonality that captures the frequent peaks and troughs through time, indicating there is a consistent annual cycle for FOODS.

The remainder component at the bottom panel, shows some short-term effects probably caused by factors such as promotions or holidays, that are not being explained by the overall trend or seasonal component. Based on these observations and the scales of the y-axis for the trend

and seasonal components, we see that for the FOODS category, a model that would adequately explain both these components would be enough to create a fairly good prediction for future demand.

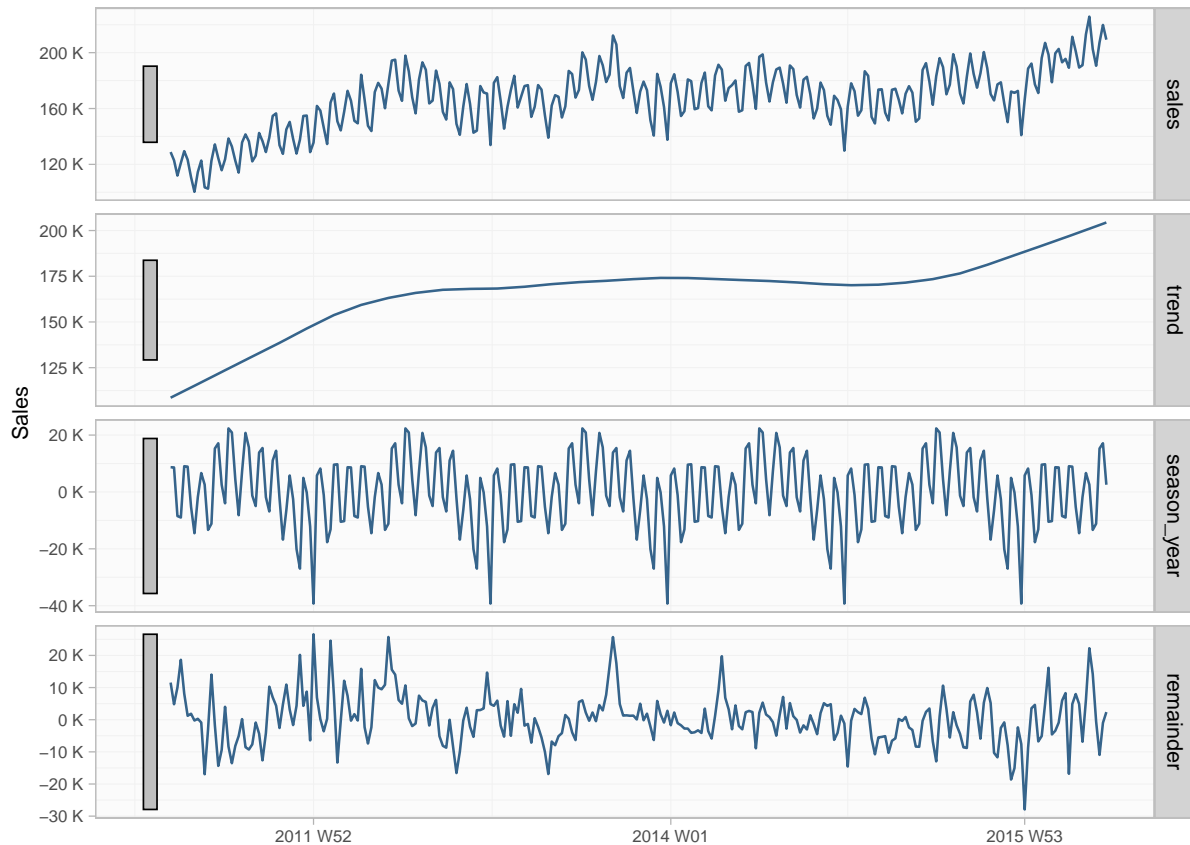


Figure 5.5.: Time series components of the FOODS category

Moving to the second decomposition group of plots for the HOBBIES series in Figure 5.6, the trend component shows a gradual upward movement, becoming more prominent from mid 2012 onward. While there is an overall increase, the trend is not as smooth or steep as that observed in FOODS aligned with the level disruptions we have highlighted earlier. The seasonal component shows the annual cycles with rather noisy peaks and troughs, indication that the model fails to locate a clear pattern. In addition, the magnitude of seasonality is much weaker compared to FOODS, indicating that demand for HOBBIES is less tied to recurring annual cycles. This aligns with the idea that this category may be more driven by irregular factors like supply disruptions. The residuals reveal significant irregularities and large spikes, particularly in the earlier years. These deviations suggest that external shocks or other unexplained factors play a substantial role for HOBBIES. The variance of the residuals decreases slightly in later years, implying some stabilization in demand after the periods of shocks.

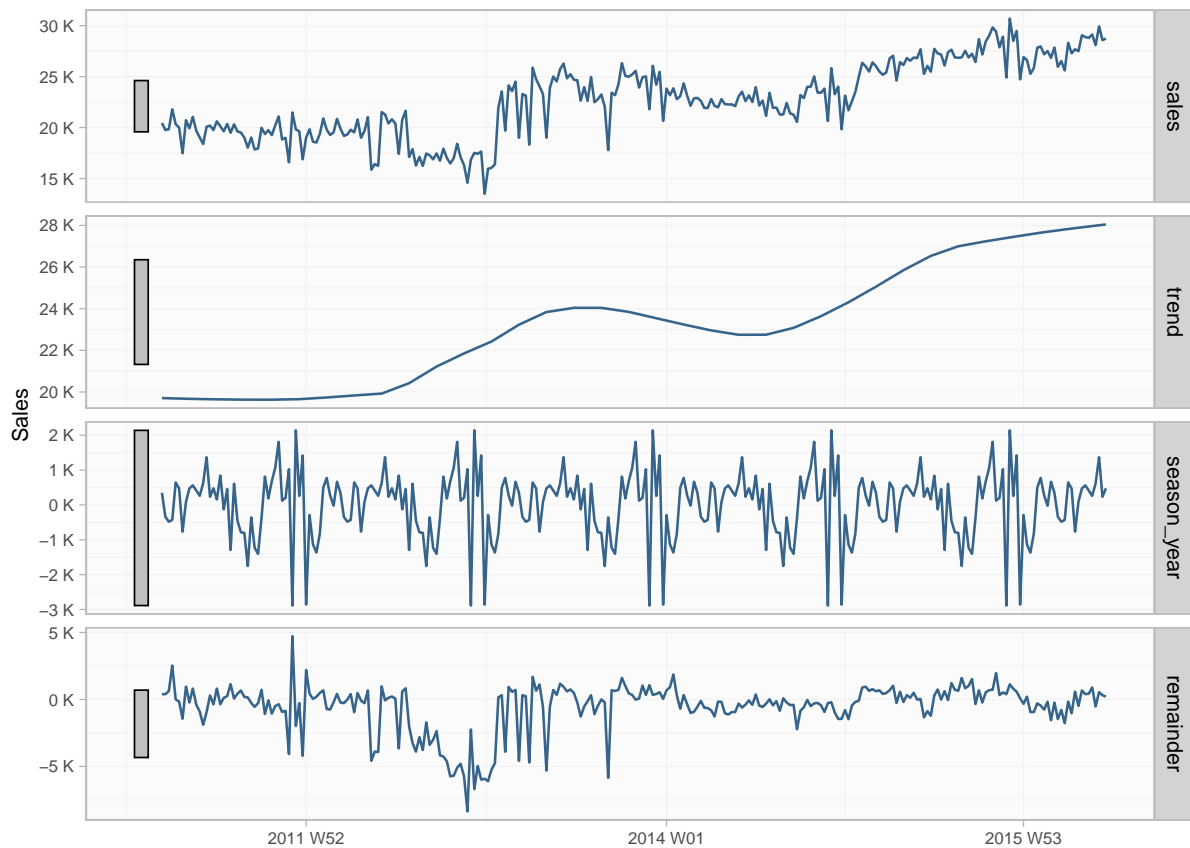


Figure 5.6.: Time series components of the HOBBIES category

For HOUSEHOLD in Figure 5.7, the trend component demonstrates a clear, gradual upward trajectory throughout the entire period. Compared to HOBBIES, the trend in this series is smoother and more consistent, with no significant slowdowns or dips, which suggests a steady increase of demand for household products over time. The seasonal component reveals a well-defined annual cycle, though the magnitude of seasonality is significantly smaller than in FOODS. Peaks and troughs are consistent across years, indicating that HOUSEHOLD products experience some level of seasonality, however, the smaller seasonal amplitude suggests that sales are less periodic-driven and more evenly distributed throughout the year compared to FOODS. The residuals are relatively stable, with fewer extreme fluctuations compared to HOBBIES. This stability implies that much of the variability in sales is captured by the trend and seasonal components. The noise in the residuals appears consistent over time, with no major structural changes, suggesting that external shocks or random events have a relatively small impact on the overall sales pattern for HOUSEHOLD. Some minor residual spikes and dips in early years (around 2012) may indicate some kind of disruptions, but these do not persist in later years.

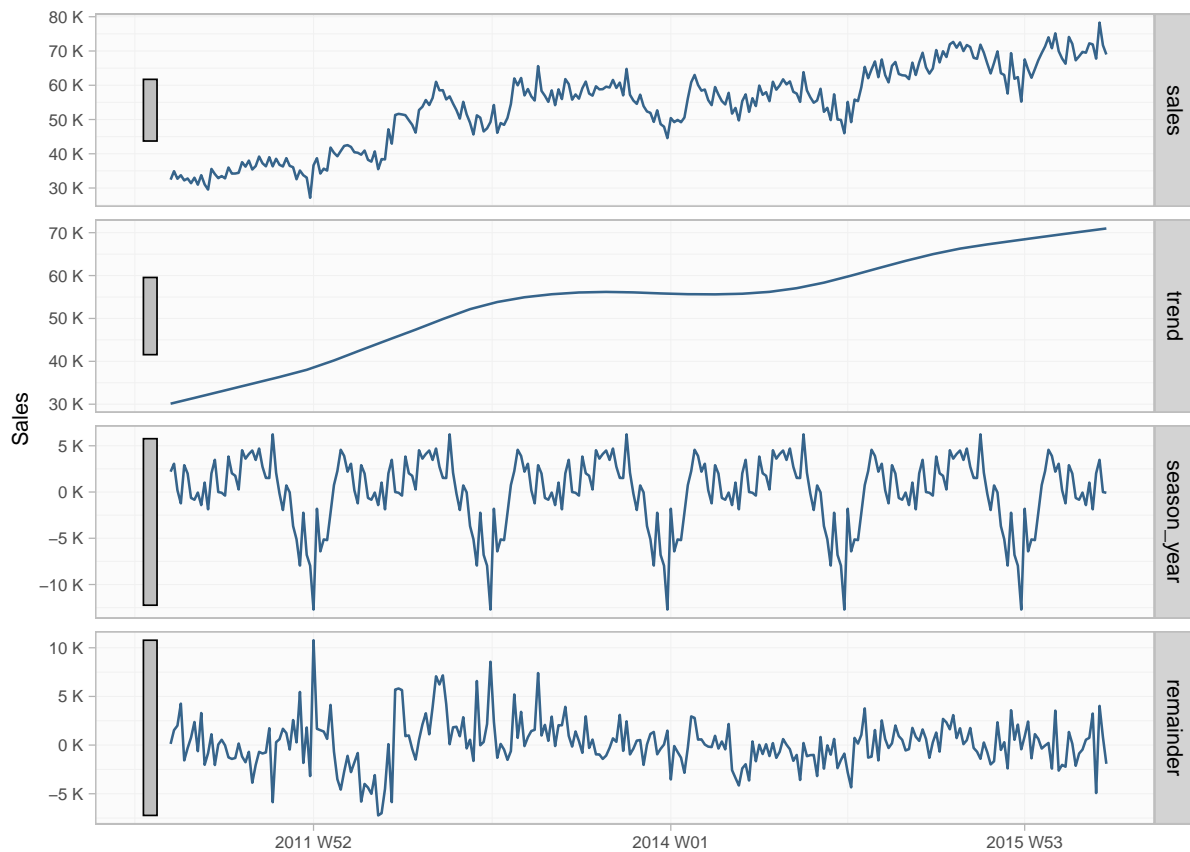


Figure 5.7.: Time series components of the HOUSEHOLD category

Overall the decomposition results have confirmed once more the insights we got from the previous parts, highlighting the importance of the trend and seasonal components for explaining the variability in FOODS and HOUSEHOLD time series. In contrast, the HOBBIES category presents a more complex case, where a significant portion of the variability cannot be attributed only from the time series components. For this category, the inclusion of other variables from the dataset might be more valuable than in the other two categories.

5.4. Autocorrelation and Lags Analysis

Given the insights we have concluded so far in our analysis, the two most significant characteristics we have are arguably the trend and the annual seasonality. In this section with the help of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), we will try to assess how strong is this trend and confirm whether the seasonality observed in decomposition persists in autocorrelations.

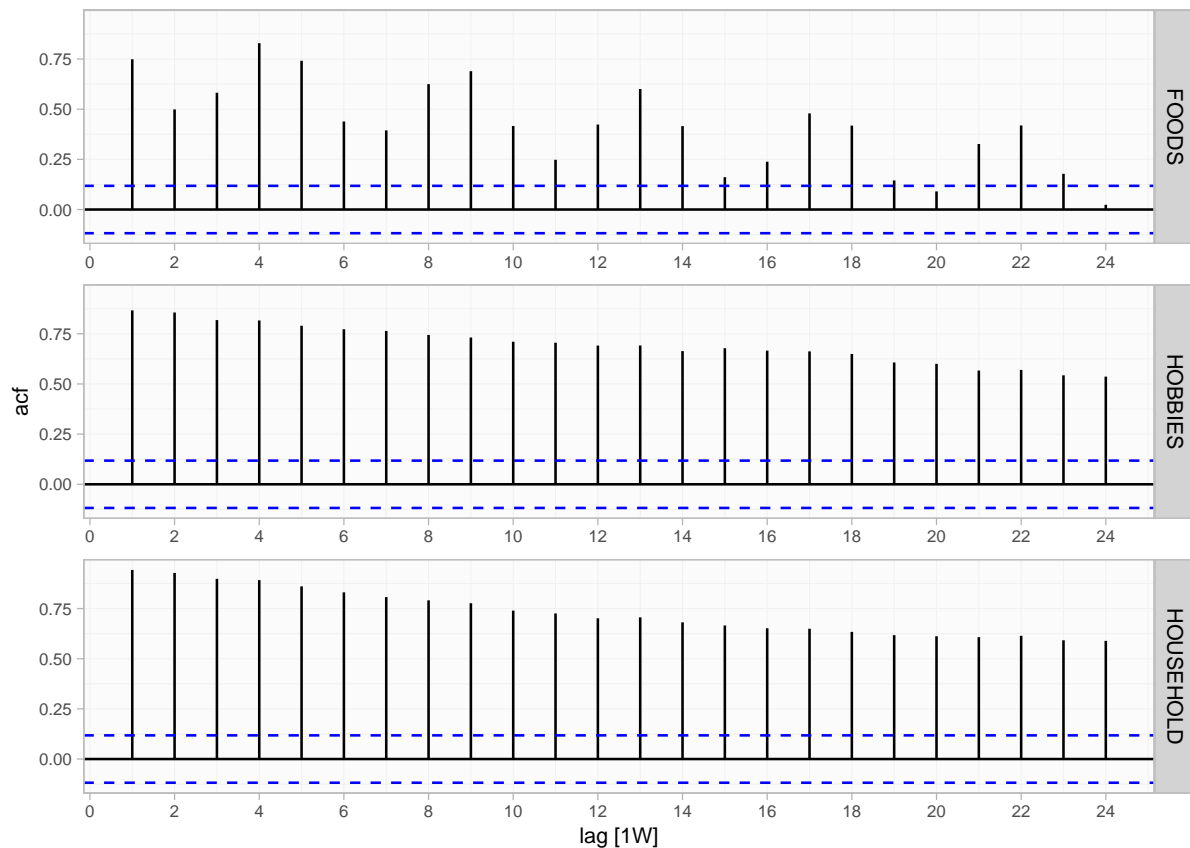


Figure 5.8.: Autocorrelation plots of sales by category

In the upper panel of Figure 5.8, the ACF plot for FOODS reveals a slow decay, indicating some non-stationary behavior, likely driven by the underlying trend component. While this aligns with prior insights, the plot shows that significant spikes persist up to the 24th lag before falling within the confidence intervals. This long persistence and the scale of the lags is a clear sign of a strong trend for this series. Additionally, the plot highlights a strong dependence of current sales on recent weeks, reflecting consistent temporal persistence in the data, while the scalloped shape with regular spikes at specific intervals (e.g., lags 4, 8, 9, 12) suggest a recurring week-of-month seasonal pattern, confirming the seasonal behavior observed in earlier analysis. Moving on to HOBBIES in the middle panel of Figure 5.8, we see that in this case too, autocorrelation is strong at short lags, with the same gradual decay we observed in FOODS. A significant difference here is that the seasonality in this series is not as apparent as was for FOODS. A closer look at the lags seasonal behavior indicates some small pattern but in any case it is weaker and potentially less consistent across time. Similar to HOBBIES, the ACF plot of HOUSEHOLD at the bottom panel, demonstrates strong dependence of current sales on recent weeks. There is high autocorrelation at short lags and a slow decay pattern as well. This aligns

with the other two series, where all maintain a significant autocorrelation over longer range of lags. Similar to HOBBIES we do not see a distinctive week-of-month seasonal pattern from this first 24 lags but the overall behavior suggests a non-stationary time series as well.

Adding to the above insights, we will use the PACF to examine the specific lags that have significant effect on each series. From the PACF plot in Figure 5.9, we see that in the FOODS series we have a more complex behavior with strong and almost equal dependencies for lags 1, 3 and 4 while there is also an inverse relationship at lags 5 and 6, probably due to seasonal behavior.

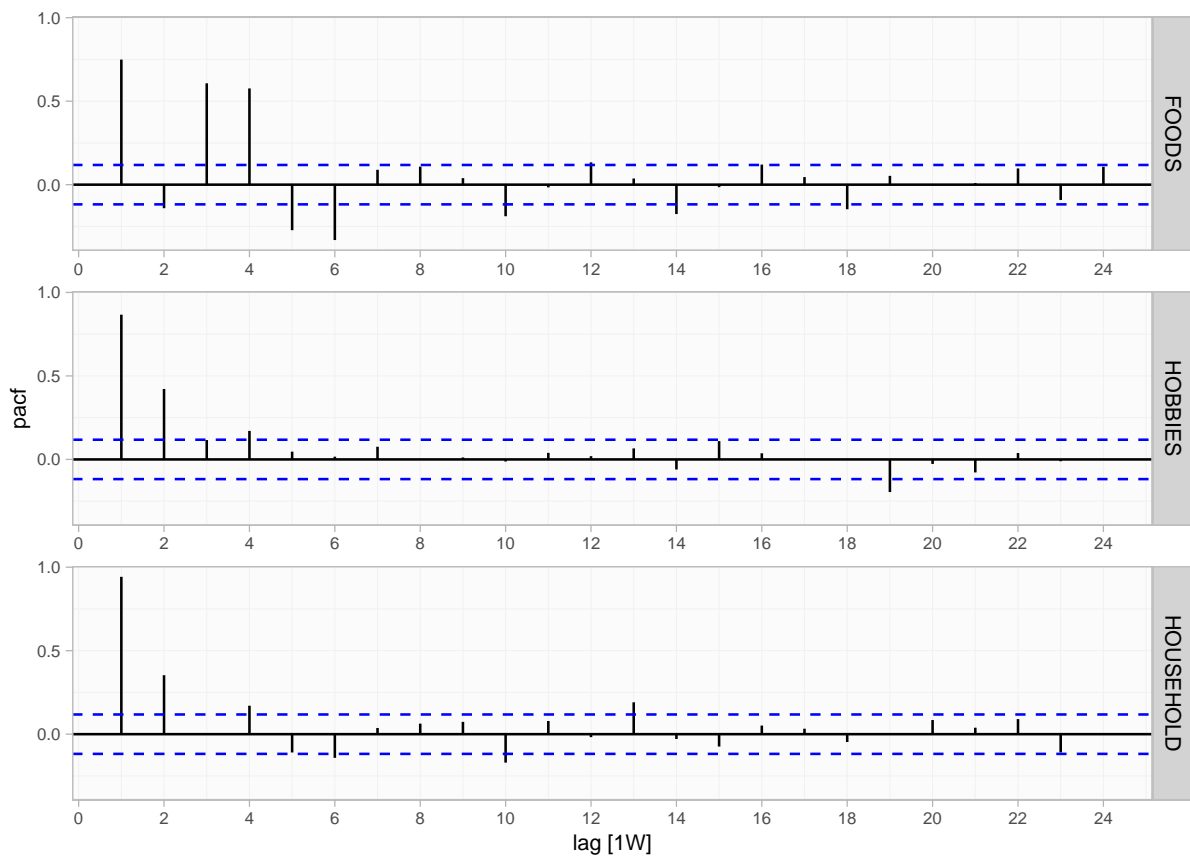


Figure 5.9.: Partial autocorrelation plots of sales by category

Following, for the HOBBIES and the HOUSEHOLD series, we observe a similar pattern, with a relatively simple structure this time, showcasing a strong relationship for lag 1 and weaker influences for lags 3 and 4.

To further investigate the significance of specific lags, we will generate plots showing the relationship between sales and their lagged values, up to lag 12. In Figure 5.10, we can see the relationships for the FOODS series.

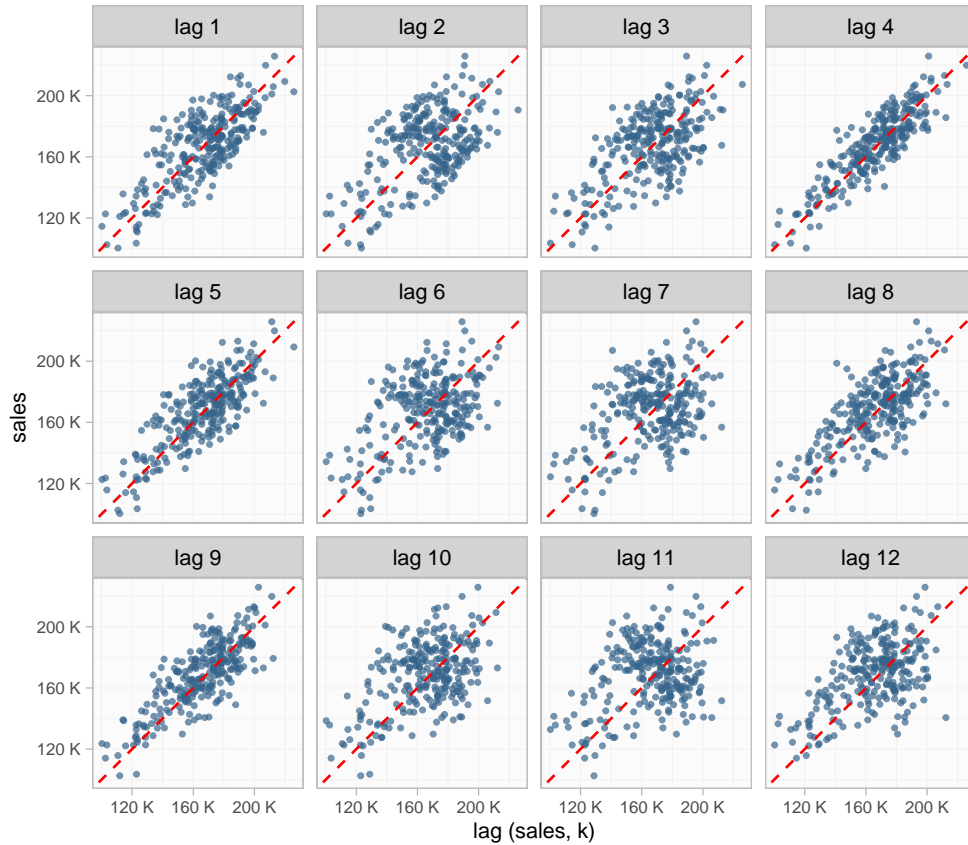


Figure 5.10.: Lag plots of weekly FOODS sales

From the visualization we can infer a strong relationship for lags 1 and 4 where we see the points are closely clustered around the line. This behavior aligns with the week-of-month cycle we have previously mentioned.

Moving on to the HOBBIES group of lag plots in Figure 5.11, we see a strong and consistent clustering of the points for most of the lags, confirming the persistent relationships identified in the ACF plot. Notably, lags 1 to 4 have the highest concentration, while the patterns become more dispersed as the lag number increases.

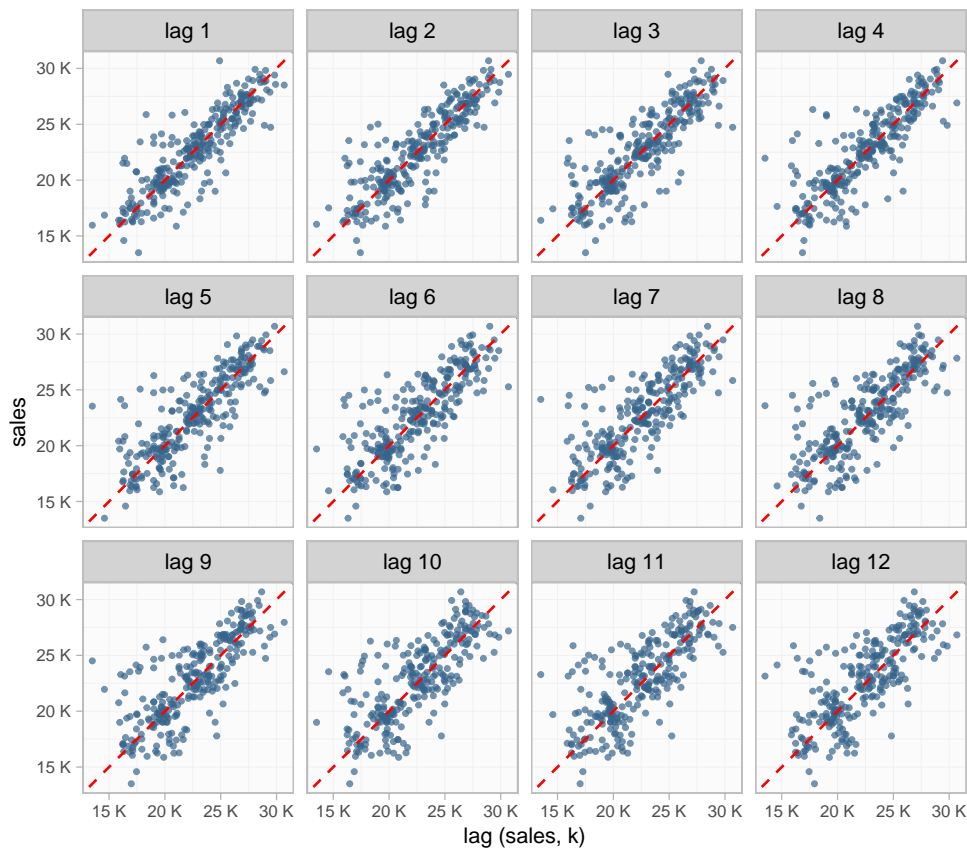


Figure 5.11.: Lag plots of weekly HOBBIES sales

Finally, for the HOUSEHOLD series lags in Figure 5.12, we seem to have similar behavior like in HOBBIES, where the concentration of the points is persistent for several lags. This persistent relationship confirms the strong autocorrelation observed in the ACF analysis, mainly up to lag 4. As the lag increases, the clustering of points becomes less tight and the relationship weakens slightly. This is expected, as the influence of past sales diminishes over time and more variability is introduced.

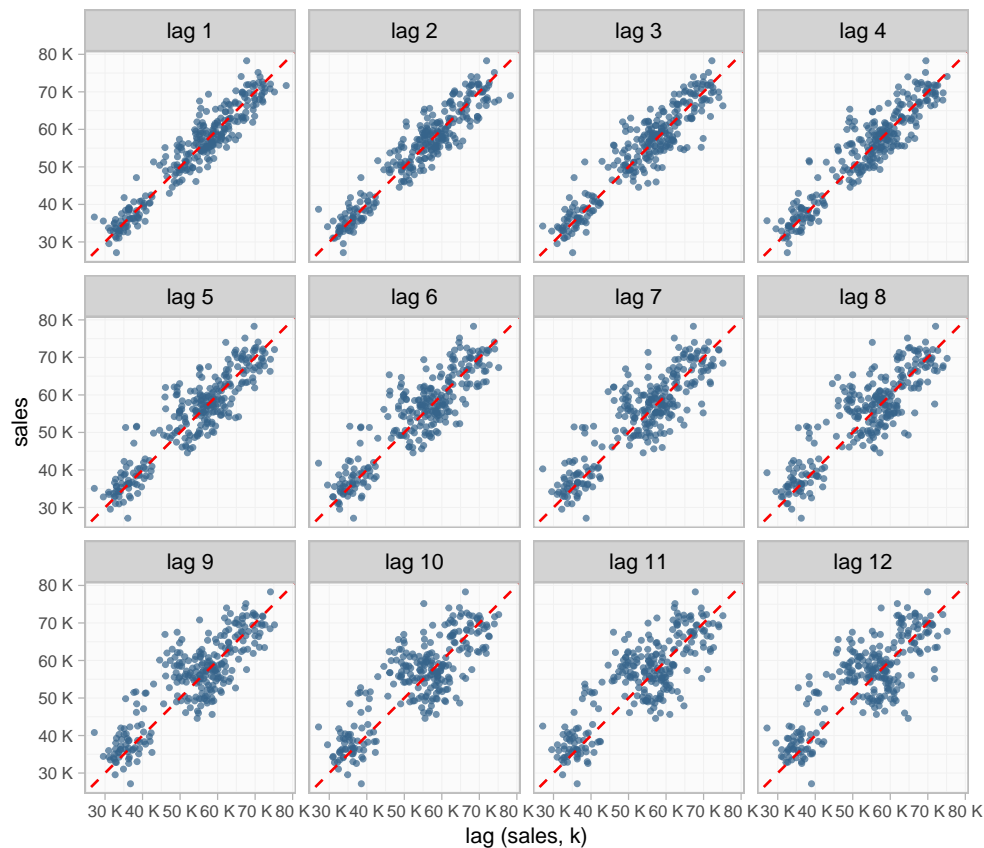


Figure 5.12.: Lag plots of weekly HOUSEHOLD sales

Overall the autocorrelation and lag analysis provided valuable insights into the temporal relationships within the sales data, highlighting the distinct characteristics of each category.

5.5. Stationarity Tests

To add to the above results, statistical stationarity tests will be applied on each product category. The already observed trending and cyclical behavior suggests non-stationarity in all three series but we need to quantify this behavior by using established tests like the Augmented Dickey-Fuller (ADF) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests.

Reading the results in Table 5.1, starting from FOODS, the p-value for the ADF test is 0.09 meaning that for a significance level of 0.05 we fail to reject the null hypothesis, suggesting that our series has a unit root and thus it is non-stationary. KPSS p-value is less than 0.01, leading to a rejection of the null hypothesis of stationarity, suggesting again that the series is non-stationary.

Both tests point to the series being non-stationary, suggesting that the presence of a trend or unit root is driving the non-stationarity as already seen.

	FOODS		HOBBIES		HOUSEHOLD	
Test	Statistic	P-value	Statistic	P-value	Statistic	P-value
KPSS	0.440	< 0.01	0.230	< 0.01	0.369	< 0.01
ADF	-3.175	0.0924	-2.849	0.2184	-3.672	0.0266

Table 5.1.: Results of stationarity tests

Similarly, both tests agree that the HOBBIES series is not stationary with a p-value less than 0.05 for KPSS and a 0.22 p-value for the ADF test. Now, for the HOUSEHOLD series results are mixed. The KPSS test results suggest the series is not stationary (p-value < 0.01) while the ADF has a borderline p-value of 0.02 suggesting that the series is stationary. This discrepancy could arise from differences in how the tests detect stationarity but combined with the previous results from visual inspection, decomposition and autocorrelation, we can safely conclude the non-stationary behavior of this series too.

5.6. Categorical Variables Over Time

Before we wrap up the time series analysis section, we will explore how the categorical variables relate to sales over time by visualizing them over time in an attempt to locate interesting behavior that could provide more information.

Starting with Figure 5.13, we have a plot depicting the weekly sales over time by category. The color scheme is on a scale from neutral grey for zero SNAP days to red for seven out of seven SNAP days in a week. The colors hint some interesting patterns especially in the FOODS panel, where weeks with a higher number of SNAP days coincide with higher sales, hinting that the benefits program positively correlates with FOODS sales performance. The other two categories however, are less conclusive than FOODS. It appears that SNAP days affect in some smaller extent the sales of HOUSEHOLD products but it does not have the same impact on HOBBIES. Nonetheless, it is reasonable to assume that the store traffic driven by SNAP benefits may indirectly contribute to sales in both the HOBBIES and HOUSEHOLD categories.

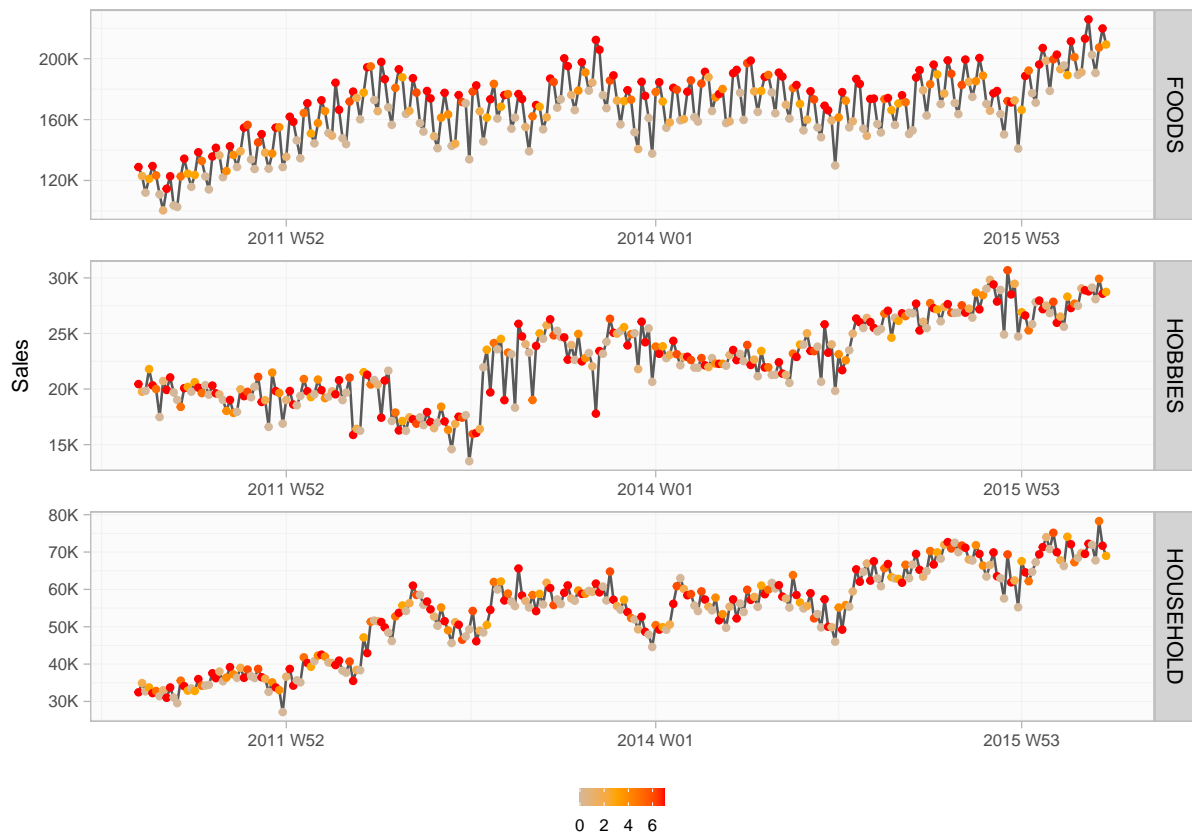


Figure 5.13.: Weekly sales with points color-coded by number of SNAP days in a week

Moving to the week-of-month cycle, we will create a new variable to mark the last two weeks of each month in order to distinguish them from the first two, effectively splitting each month into starting and ending weeks. This will allow us to investigate whether sales tend to decline as the month progresses. As shown in Figure 5.14 FOODS sales exhibit a clear pattern where ending weeks (red points) tend to occur in troughs, indicating lower sales compared to starting weeks (green points). This trend, likely associated with payment cycles, is noticeable stronger in FOODS than in the other categories. HOUSEHOLD shows a moderate effect while HOBBIES appears relatively unaffected.

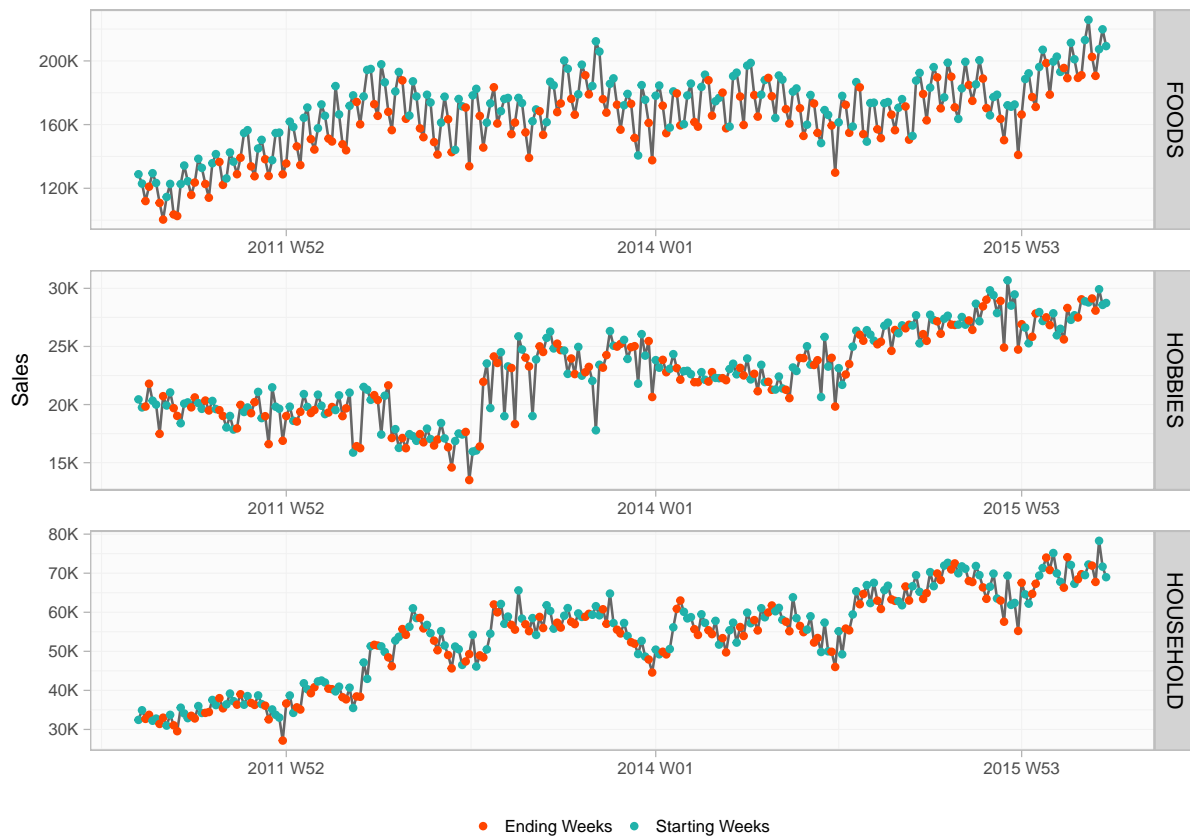


Figure 5.14.: Impact of the last week of month on sales over time

5.7. Key Takeaways

The findings from this chapter highlight the strong influence of the trend component in our series, as well as the presence of multiple seasonalities (annual and week-of-month cycles). The strong autocorrelation is another important feature, which we might need to mitigate in the modeling phase utilizing the lags we have identified as significant.

We are now moving to the next chapter, Regression Analysis, where we will systematically test the statistical significance of the insights gathered. The results of these models will enable an evaluation of the explanatory power of each factor and help us answer the main research questions of this research.

6. Regression Analysis

6.1. Modeling Process

In this chapter, regression analysis will be used for a systematic investigation of the explanatory factors on sales. Based on the exploratory and time series analysis carried out in the last two chapters, here we focus on identifying the drivers of the sales trends for each category. Multiple linear regression will be applied to study the statistical significance of these factors while also ensuring that the assumptions of the OLS are met and results are valid.

The modeling process is structured to ensure both robustness and reliability. Initial modeling is done using variables which were found from earlier analysis, such as the weekly mean price, holidays and SNAP days. The statistical performance of each model is checked through statistical measures such as adjusted R^2 and p-values while multicollinearity is assessed through VIF. The initial models go through a series of diagnostic checks to test the assumptions of regression analysis (linearity, homoscedasticity and independence of residuals). KPSS and ADF tests are applied to ensure stationarity of residuals and the Ljung–Box test is applied in order to assess their independence. Additionally, the Breusch-Pagan test checks the constant variance of residuals while the Shapiro-Wilk test is used to check their normality.

If the initial models do not meet the OLS assumptions, a second regression model with the addition of lagged variables is developed to capture possible dependencies and improve the model's validity. The number of lags that are added to each model have been identified in the previous chapter. This iterative approach that follows the process of the framework described in Chapter 2, allows for the refinement of models, ensuring their predictive reliability and interpretability. Through this process, the regression analysis provides insights into the factors influencing sales within each series. These findings set the stage for the evaluation of the models' predictive accuracy in Chapter 7.

6.2. Regression Analysis - FOODS

6.2.1. Initial Regression Model

In order to test the statistical properties of the FOODS sales time series and their coefficients' significance we will create the following multiple regression model:

$$\begin{aligned} \text{Sales}_{\text{FOODS},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} \\ & + \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Event} + \beta_{16} \cdot \text{Price} + \varepsilon_t \end{aligned} \quad (6.1)$$

where:

- β_0 is the intercept term (baseline) t
- Trend_t is a numeric time-trend variable capturing overall growth or decline over time
- $\sum_{m=2}^{12} \beta_m \cdot \text{Month}_m$ is the set of month dummy variables (month 1 is the omitted reference category)
- End-Week is a dummy indicator for whether a given week is at the end of a month (1 = yes)
- SNAP is a dummy indicator for weeks that include more than 2 SNAP days
- Event is a dummy indicator for whether a given week includes any events or holidays
- Price is the average price of sold products in period t
- ε_t is the residual (error) term

Before assessing the fit of the model, we will proceed with the residuals diagnostics to assess whether the model meets the regression assumptions.

6.2.2. Residuals Diagnostics

From the diagnostics plots in Figure 6.1, we see a cyclical pattern in the residuals plot at the top panel indicating that the model fails to catch the trending behavior. This behavior has been passing on into the residuals making them non-stationary. Moreover ACF plot indicates a high autocorrelation and the slight curvature in the scale-location plot suggests possible heteroscedasticity. Taken together, these issues imply that the model is inadequate, as several regression assumptions are not met and despite the residuals' normality and distribution that seems acceptable, further refinement is needed.

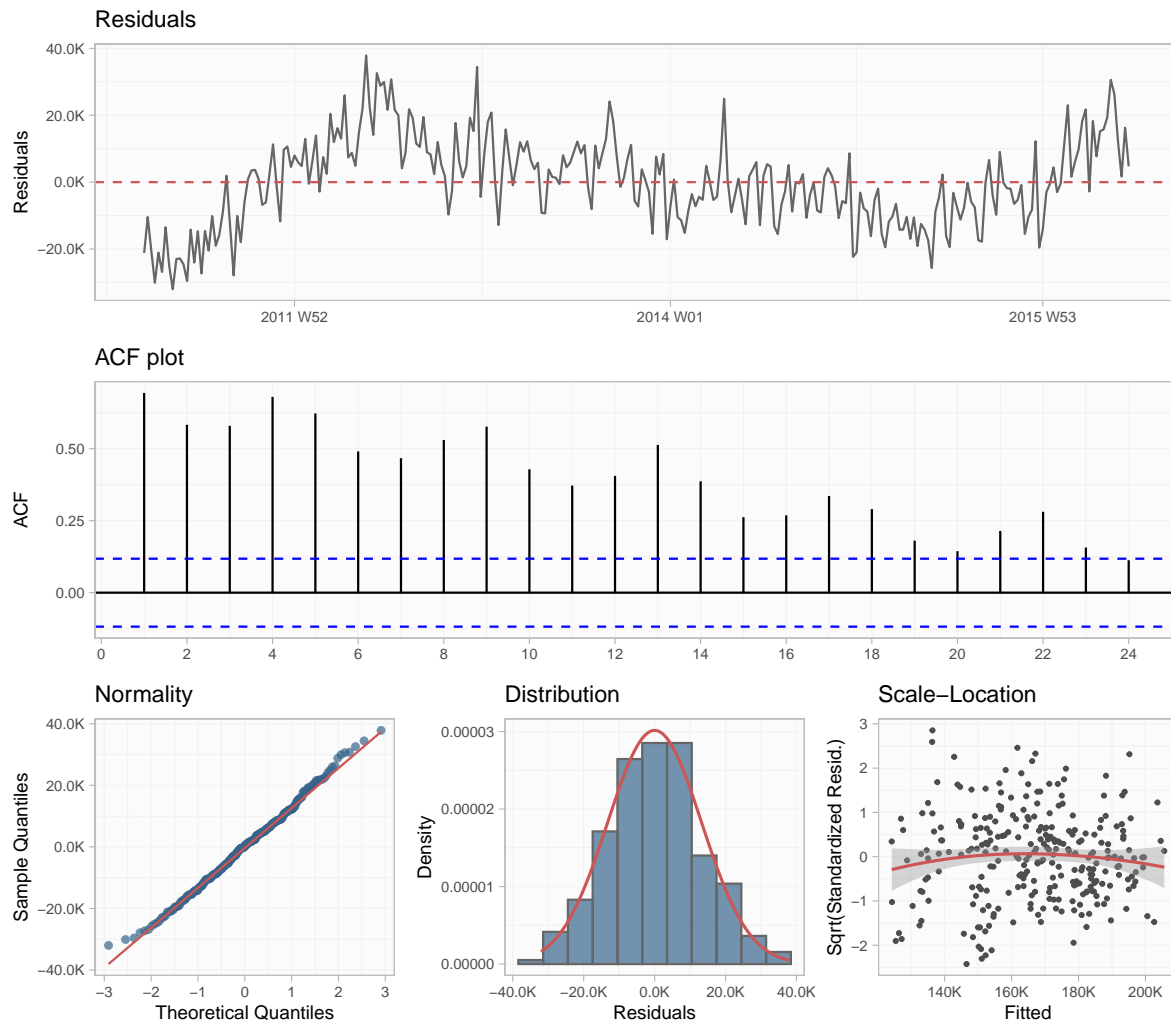


Figure 6.1.: Residuals diagnostic plots of the initial FOODS regression model fit

To quantify the above inferences we will use statistical tests to assess stationarity, autocorrelation, normality and the presence of heteroscedasticity in the models' residuals. The results in Table 6.1 confirm the visual inspection. ADF clearly shows a non-stationary behavior while KPSS gives a borderline result ($p\text{-value} = 0.065$), though considering also the visual inspection, we can conclude that the residuals are not stationary. Moving to autocorrelation testing, Ljung-Box test reports back a $p\text{-value}$ less than 0.01 (actual value is much less than 0.01), indicating that residuals exhibit autocorrelation. In the same manner, the very low $p\text{-value}$ of the Breusch-Pagan test lead us to reject the null hypothesis of homoscedasticity, indicating that the residuals exhibit heteroskedasticity. Last, the Shapiro-Wilk finds the residuals normal ($p\text{-value} = 0.717$) but given that all the other tests failed, we must move on to developing a more suitable model that meets the regression assumptions and proves robust for this analysis.

Test	Statistic	P-value
KPSS	0.428	0.065
ADF	-2.739	0.265
Ljung-Box	134.224	< 0.01
Breusch-Pagan	14.084	< 0.01
Shapiro-Wilk	0.996	0.717

Table 6.1.: Tests results on the residuals of the initial FOODS model fit

6.2.3. Refined Regression Model

The strong autocorrelation in the residuals of the initial FOODS model suggests that past values still carry information not captured by the explanatory variables we used. In order to reduce autocorrelation, stabilize variance and generally satisfy the regression assumptions we will add lagged terms of the dependent variable Sales, to explicitly account for these dynamic relationships. From the lags analysis performed in Chapter 5, we have highlighted that lag 1 and lag 4 can help us account for this short-term memory, so we will add them to the initial model:

$$\begin{aligned}
 \text{Sales}_{\text{FOODS},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} + \\
 & \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Event} + \beta_{16} \cdot \text{Price} + \beta_{17} \cdot \text{Lag 1} + \\
 & \beta_{18} \cdot \text{Lag 4} + \varepsilon_t
 \end{aligned} \tag{6.2}$$

where:

- Lag 1 is the sales value one period before t
- Lag 4 is the sales value four periods before t

6.2.4. Residual Diagnostics for the Refined Model

The residuals of this second model with the addition of the lagged predictors, resulted in a clearly improved residual behavior. From the visual inspection in Figure 6.2 we witness a rather stationary behavior of the residuals, though some variability still exists. The magnitude of the variability though, is not that significant. Moreover, the ACF plot suggests that our new model

handles much better the short memory of the data, without significant peaks outside the significance bounds indicating that residuals are white noise. At the bottom, the three panels show satisfactory results but we need to confirm the visual inspection with the statistical tests once more.

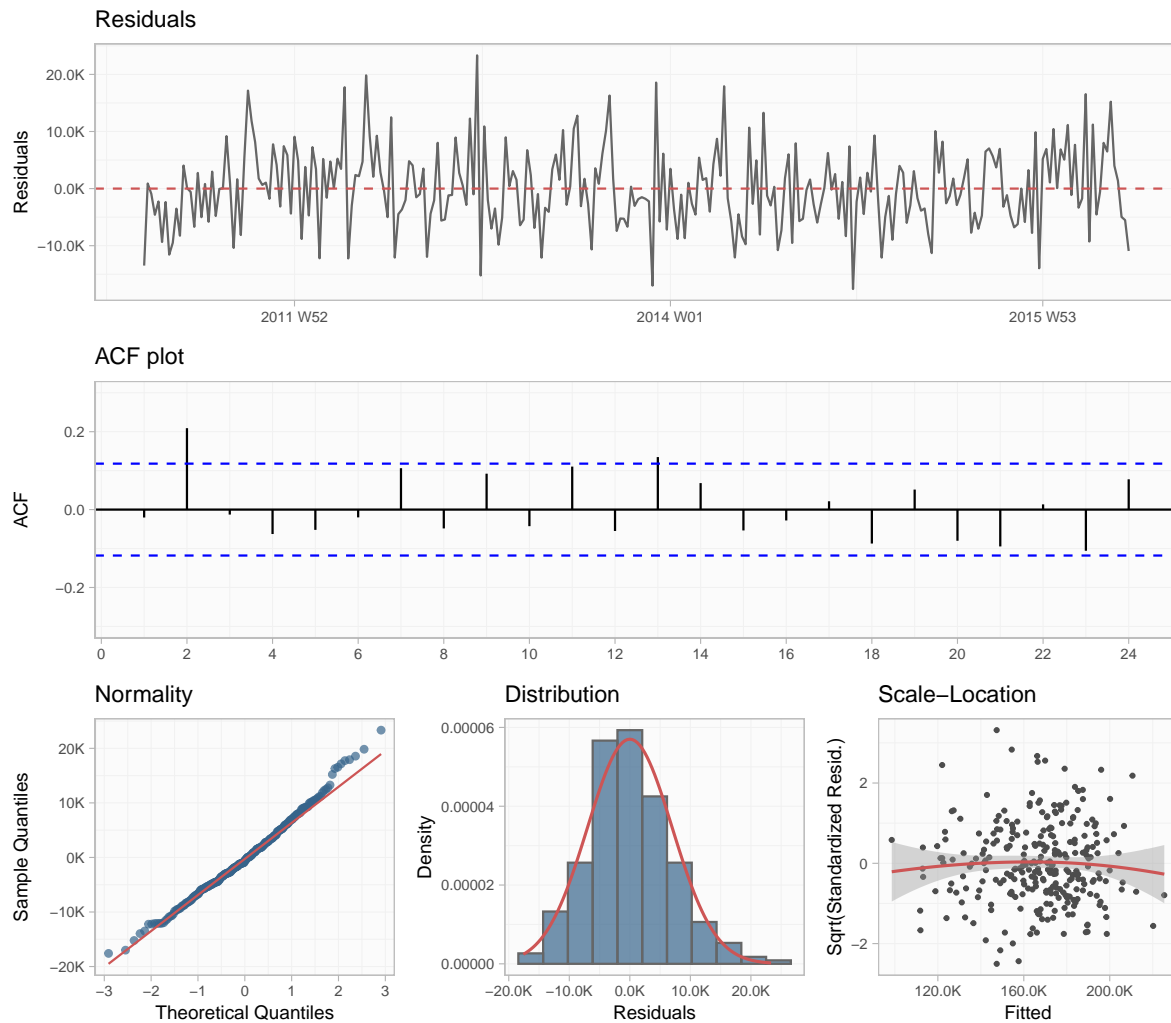


Figure 6.2.: Residuals diagnostics plots of the refined FOODS regression model fit

The results of the tests applied on the residuals of the refined FOODS regression model (Table 6.2) confirm the visual inspection. Stationarity tests both find the residuals stationary (no unit root), while the Ljung-Box test p-value (0.7341) is well above the significance threshold of 0.05., so we fail to reject the null hypothesis of no autocorrelation. Finally the Breusch-Pagan indicates no presence of heteroscedasticity and the normality test on the residuals is satisfactory.

Test	Statistic	P-value
KPSS	0.118	0.1
ADF	-5.592	< 0.01
Ljung-Box	0.115	0.7341
Breusch-Pagan	0.000	0.9885
Shapiro-Wilk	0.991	0.0858

Table 6.2.: Tests results on the residuals of the refined FOODS model fit

Overall we have a robust model, much improved by the addition of the lagged variables, so now we can move forward to the assessment of the goodness-of-fit for this refined model.

6.2.5. Fit Assessment

The result statistics of the fit of the refined FOODS regression model can be seen in Table 6.3. The model's R^2 and adjusted R^2 imply that around 90% of the variability in the dependent variable is explained by the regression, which is an impressive explanatory power. The difference between R^2 and adjusted R^2 is small, suggesting that the model is not overfitting while the overall p-value is less than 0.001, so the model is considered statistically significant.

Regarding our key predictors, Trend variable has a positive and significant coefficient implying that, on average, sales are slightly rising over time. This is consistent with our inferences from the exploratory sections. Moving on to the Month dummies, we see that most months are found statistically significant, while the End-Week dummy shows some evidence that weeks of the second half of each month might show reduced sales, but this effect is near the edge of the 0.05 significance level. This can be attributed to the addition of the lagged terms that may have reduced the significance of this variable. SNAP days dummy has a highly significant result with large positive coefficient, suggesting that on SNAP days, sales are higher by nearly 10,000 units compared to non-SNAP days. This fact confirms once more, the assumptions we have made during our preliminary data explorations.

Next, the Event dummy variable is not found statistically significant, suggesting events as defined here do not systematically raise or lower the weekly sales in a reliable way. Regarding the Price variable, a negative coefficient (-6,442) aligns with our expectations that higher prices depress sales, but the effect of this variable is found not statistically significant in this model.

Refined FOODS Model				
Variable	Estimate	95% CI	P-value	GVIF
(Intercept)	32,234	-12,476, 76,943	0.157	
Trend	24	0.47, 47	0.046	4.7
Month dummies				1.5
<i>Jan</i>	—	—		
<i>Feb</i>	-2,553	-7,124, 2,017	0.272	
<i>Mar</i>	-5,908	-10,142, -1,674	0.006	
<i>Apr</i>	-5,131	-9,515, -747	0.022	
<i>May</i>	-4,870	-9,266, -474	0.030	
<i>Jun</i>	-406	-4,626, 3,814	0.850	
<i>Jul</i>	-3,410	-8,073, 1,252	0.151	
<i>Aug</i>	-3,058	-7,775, 1,660	0.203	
<i>Sept</i>	-5,014	-9,523, -506	0.029	
<i>Oct</i>	-8,380	-13,042, -3,718	<0.001	
<i>Nov</i>	-9,530	-14,069, -4,991	<0.001	
<i>Dec</i>	-7,227	-11,566, -2,887	0.001	
End-Week dummy				2.7
0	—	—		
1	-2,578	-5,402, 245	0.073	
SNAP dummy				2.3
0	—	—		
1	9,932	7,332, 12,532	<0.001	
Event dummy				1.2
0	—	—		
1	1,123	-742, 2,988	0.237	
Price	-6,442	-20,462, 7,579	0.366	2.4
Lag 1	0.44	0.38, 0.50	<0.001	2.8
Lag 4	0.46	0.38, 0.54	<0.001	4.6
<i>R</i> ²	0.909			
Adjusted <i>R</i> ²	0.902			
Statistic	142			
<i>df</i>	18			
<i>p</i> -value	<0.001			

Abbreviations: CI = Confidence Interval, GVIF = Generalized Variance Inflation Factor

Table 6.3.: Summary of the refined regression model for the FOODS time series

Finally, the additional Lags of sales are both found strongly significant and positive indicating a degree of persistence in sales where higher sales in the recent past predict higher sales in the current period. The table also lists GVIF values, with most in the 1–2.7 range and the highest around 4.7 for the Trend term. Since all of the variables are below 10, there is not much concern regarding multicollinearity issues.

Following the above statistical assessment, in Figure 6.3, we see the fitted values plotted against the actual sales of the FOODS category. The plot, aligned with the above results, demonstrates an impressive fit, with the predicted values from the model, closely following the seasonal variance and the overall trend, though missing on some extreme values. The effect of the lagged

variables is evident from the fact that the fitted line is successfully adapting to sudden changes, for example at the beginning of 2016, when trend becomes more steep.

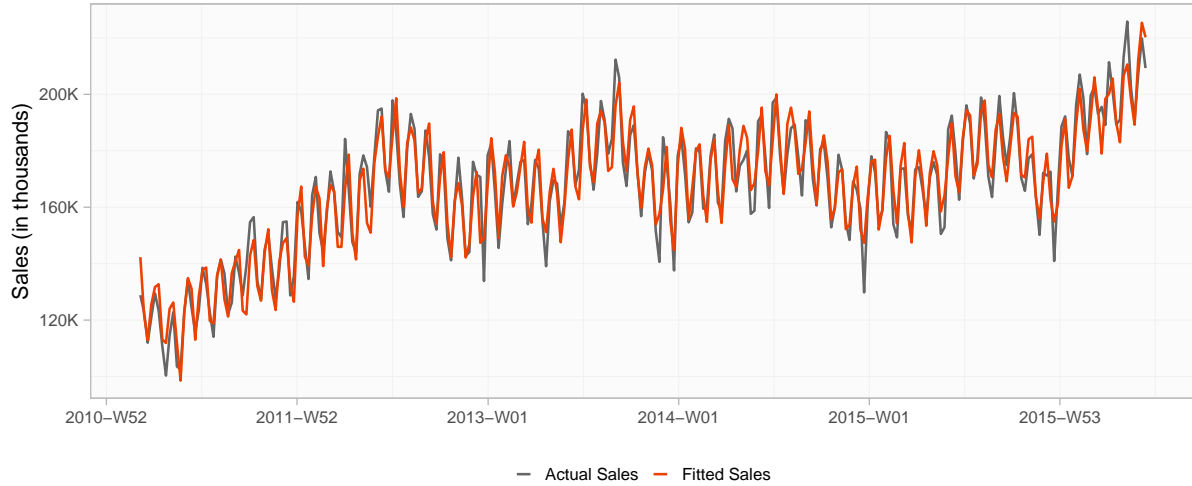


Figure 6.3.: Plot of the refined FOODS model fit vs original values

Overall, this regression effectively explains the variation in sales, especially through seasonality, SNAP effects and autocorrelation (sales lags). The model thus appears to be a good fit for the data, with only a few predictors (Events, Price) not showing significance.

6.3. Regression Analysis - HOBBIES

6.3.1. Initial Regression Model

For this next part of regression analysis we will built the initial model for the HOOBIES category using the baseline equation:

$$\begin{aligned} \text{Sales}_{\text{HOBBIES},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} \\ & + \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Event} + \beta_{16} \cdot \text{Price} + \varepsilon_t \end{aligned}$$

After specifying the components of the HOBBIES model we will proceed with the evaluation of the residuals to ensure they meet the assumptions of regression modeling.

6.3.2. Residual Diagnostics

Starting from the residuals plot at the top panel in Figure 6.4, we see the residuals fluctuating around zero, which is desired, but the wavy pattern indicates that there is information passed into residuals that the model could not efficiently capture. Once more the ACF plot indicates that residuals are not independent at multiple lags while Q-Q plot deviates from normality especially in the tails and distribution exhibits a slight skewness. As inferred from the amplitude of residuals we have already observed, the scale-location plot shows a curved trend proving that variance of the residuals is not constant.

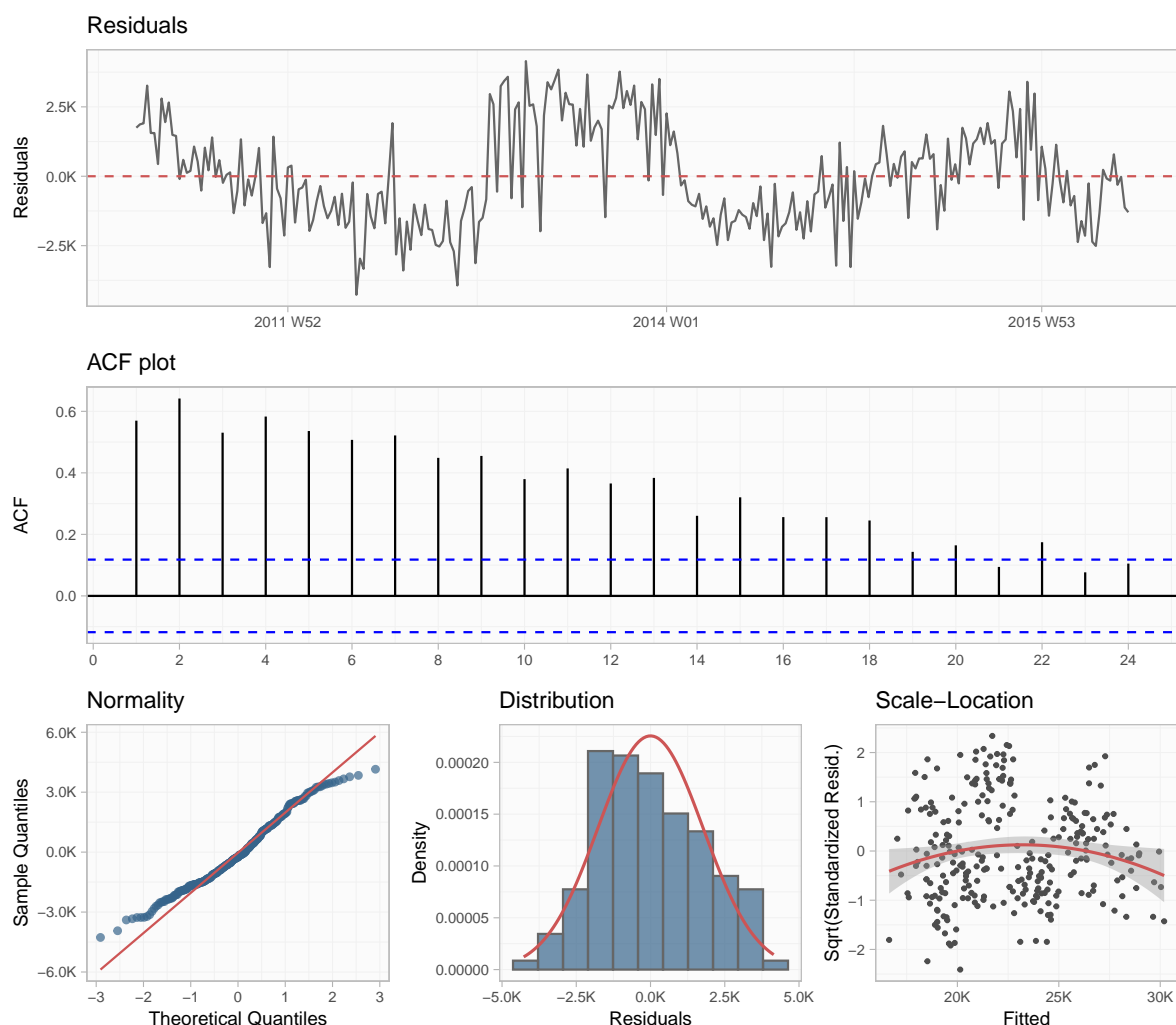


Figure 6.4.: Residuals diagnostic plots of the initial HOBBIES regression model fit

The tests results come in accordance to the visuals, where stationarity is not conclusive from the ADF and KPSS tests, while the Ljung-Box as expected, does not fail to reject the null hypothesis

of no autocorrelation. Heteroscedasticity is confirmed by the low p-value in the Breusch-Pagan test and finally normality of the residuals is not met, with the Shapiro-Wilk p-value being much lower than the 0.05 significance level.

Test	Statistic	P-value
KPSS	0.190	0.1
ADF	-2.321	0.4406
Ljung-Box	90.536	< 0.01
Breusch-Pagan	8.992	< 0.01
Shapiro-Wilk	0.981	< 0.01

Table 6.4.: Tests results on the residuals of the initial HOBBIES model fit

Given the above results, this time series will be challenging to model, thus we will proceed with adding further explanatory variables to get a robust model.

6.3.3. Refined Regression Model

Due to the overall behavior of the HOBBIES time series and the inadequacy of the initial model to capture the variability in the data we will introduce lagged predictors of the sales in order to refine our model. The new model's structure, including the lags we found significant in Chapter 5, will be:

$$\begin{aligned}
 \text{Sales}_{\text{HOBBIES},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} + \\
 & \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Event} + \beta_{16} \cdot \text{Price} + \beta_{17} \cdot \text{Lag 1} + \\
 & \beta_{18} \cdot \text{Lag 2} + \beta_{19} \cdot \text{Lag 3} + \varepsilon_t
 \end{aligned} \tag{6.3}$$

where:

- Lag 1 is the sales value one period before t
- Lag 2 is the sales value two periods before t
- Lag 3 is the sales value three periods before t

6.3.4. Residual Diagnostics for the Refined Model

The newly defined model shows an improved residuals behavior. In Figure 6.5, we see residuals appear to be relatively consistent around the zero line, suggesting that the model does not systematically overestimate or underestimate sales. We also witness an increased variability during the second half of the year 2012 and the first half of year 2013, where we have the structural changes in the HOBBIES time series. Due to the lack of other explanatory variables, these structural changes are found difficult to model and thus, we see this behavior in the residuals.

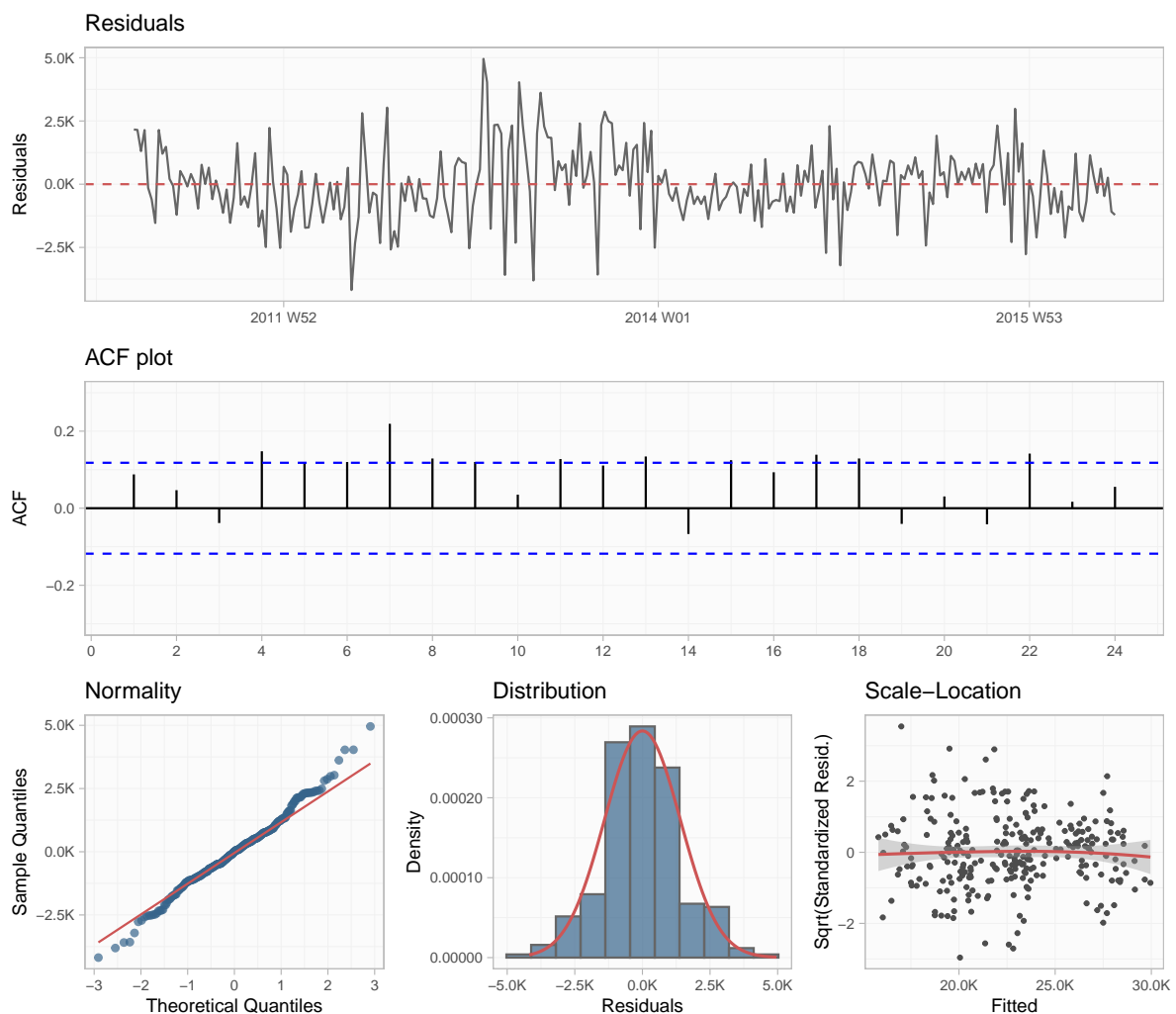


Figure 6.5.: Residuals diagnostic plots of the refined HOBBIES regression model fit

Apart from that, the ACF plot indicates no significant autocorrelation in the residuals, with only some random spikes exceeding the confidence bounds. The residuals are approximately normally distributed, with some minor issues - probably caused by the structural changes - observed

at the extremes and a roughly symmetrical distribution. Regarding the Q-Q plot we see that the points align well with the 45-degree line, with only minor deviations at the extremes suggesting residuals are approximately normally distributed, which is a significant improvement compared to the initial model, while the histogram gives back a roughly symmetric distribution, indicating that normality is not a major concern. The Scale-Location plot shows a relatively flat red trend line, indicating that the variance of residuals does not change significantly, so there is no clear evidence of heteroskedasticity.

Moving on to quantify the residual behavior of the refined model, we see that both the ADF and the KPSS tests meet the stationarity assumption and the Ljung-Box confirms that autocorrelation is no longer a concern.

Test	Statistic	P-value
KPSS	0.191	0.1
ADF	-3.690	0.0249
Ljung-Box	2.151	0.1425
Breusch-Pagan	6.212	0.0127
Shapiro-Wilk	0.990	0.059

Table 6.5.: Tests results on the residuals of the refined HOBBIES model fit

Interestingly, the Breusch-Pagan test is still detecting potential heteroscedasticity (p-value 0.0127). This is likely due to the design of this test, which makes it particularly sensitive to structural changes in the data and results in detecting these localized variance shifts that we already mentioned. However, since the scale-location plot shows no systematic pattern of change in variance and all the other residual diagnostics confirm that the model is well-specified, we will proceed with it as is.

6.3.5. Fit Assessment

The challenging nature of the HOBBIES time series is evident when we look at the results of the fit. In Table 6.6 we see that we have a statistically significant model that successfully explains 85% of the overall variation in the dependent variable ($R^2 = 0.85$, adjusted $R^2 = 0.84$). However, the majority of the key predictors are found not significant. Trend is reported as strongly significant variable, confirming that HOBBIES sales are increasing over time by about 20 units per period, on average. However we see a GVIF value of 9.7, which is somewhat high but in any case this does not invalidate the model since this inflated GVIF value is caused from the introduction of the three lagged variables in the model. Next, all months compared to the

baseline (January) show non-significant differences and overall, there does not appear to be a strong seasonal pattern for HOBBIES. The same applies for the other three dummy variables in the model, where none of which is found statistically significant. On the other side we have a strongly significant and negative coefficient for Price, indicating that increases in average price are associated with lower HOBBIES sales. Last, all three lagged variables have positive and significant coefficients meaning that HOBBIES sales strongly depend on recent past sales.

Refined HOBBIES Model				
Variable	Estimate	95% CI	P-value	GVIF
(Intercept)	14,966	10,160, 19,773	<0.001	
Trend	20	14, 27	<0.001	9.7
Month dummies				1.7
Jan	—	—		
Feb	467	-475, 1,409	0.330	
Mar	-52	-923, 818	0.906	
Apr	231	-677, 1,140	0.616	
May	-371	-1,278, 535	0.421	
Jun	328	-564, 1,219	0.470	
Jul	-237	-1,193, 720	0.626	
Aug	-583	-1,528, 361	0.225	
Sept	-511	-1,397, 376	0.258	
Oct	-70	-1,006, 867	0.884	
Nov	-814	-1,742, 114	0.085	
Dec	-718	-1,599, 162	0.109	
End-Week dummy				2.0
0	—	—		
1	-412	-907, 84	0.103	
SNAP dummy				2.0
0	—	—		
1	410	-80, 900	0.101	
Event dummy				1.2
0	—	—		
1	-316	-700, 68	0.106	
Price	-1,967	-2,782, -1,152	<0.001	3.9
Lag 1	0.22	0.10, 0.34	<0.001	6.2
Lag 2	0.31	0.20, 0.43	<0.001	5.7
Lag 3	0.14	0.03, 0.26	0.014	5.7
R ²	0.851			
Adjusted R ²	0.840			
Statistic	77.1			
df	19			
p-value	<0.001			

Abbreviations: CI = Confidence Interval, GVIF = Generalized Variance Inflation Factor

Table 6.6.: Summary of the refined regression model for the HOBBIES time series

Visualizing the fit of the refined HOBBIES model in Figure 6.6, we see that the model with the help of the lags, successfully tracks the overall upward trend in HOBBIES sales over time, especially after the first half of 2013. It is also clear that it struggles to match the short-term fluctuations and the magnitude of some shocks that appear between 2012 and 2014.

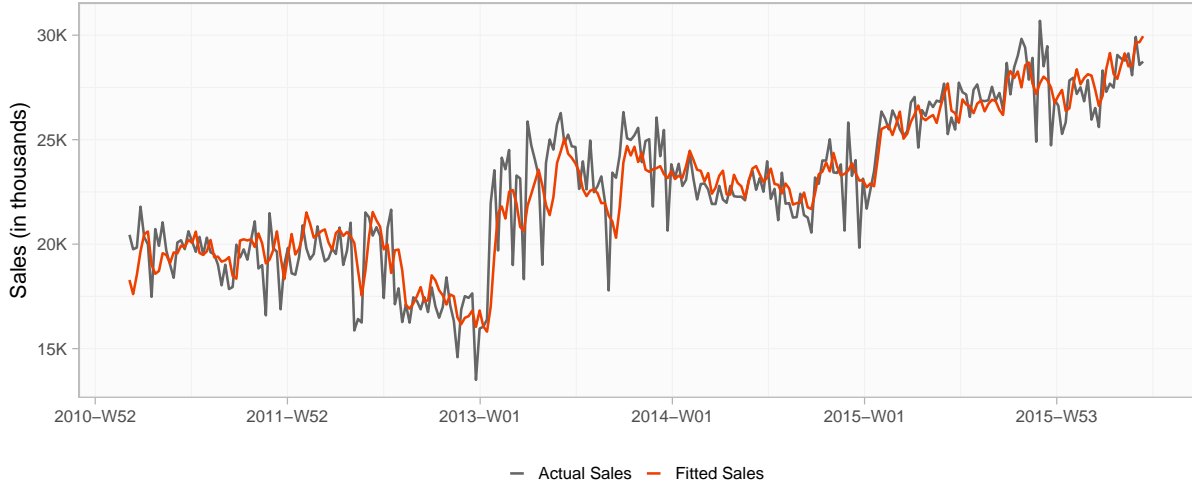


Figure 6.6.: Plot of the refined HOBBIES model fit vs original values

However the lack of explanatory variables for this irregular behavior does not compromise the overall performance, resulting in an adequate fit.

6.4. Regression Analysis - HOUSEHOLD

6.4.1. Initial Regression Model

Finally, to build the same baseline model we will apply on the HOUSEHOLD data the regression equation:

$$\text{Sales}_{\text{HSHLD},t} = \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} + \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Event} + \beta_{16} \cdot \text{Price} + \varepsilon_t \quad (6.4)$$

After the initial model is fitted on the HOUSEHOLD data, we can proceed with the residuals evaluation in the next section.

6.4.2. Residual Diagnostics

The residuals of the initial HOUSEHOLD regression model show a cyclical pattern suggesting that the model does not fully capture certain time-related effects. In addition, the ACF plot

shows significant autocorrelation at multiple lags and signs of seasonality, with peaks every 4 to 5 weeks, indicating that the residuals from this model are not independent and thus, violating another key assumption of regression. Last, at the bottom panels although the normality assumption seems to be met, in the scale-location plot we see a curvy pattern indication of heteroscedastic behavior.

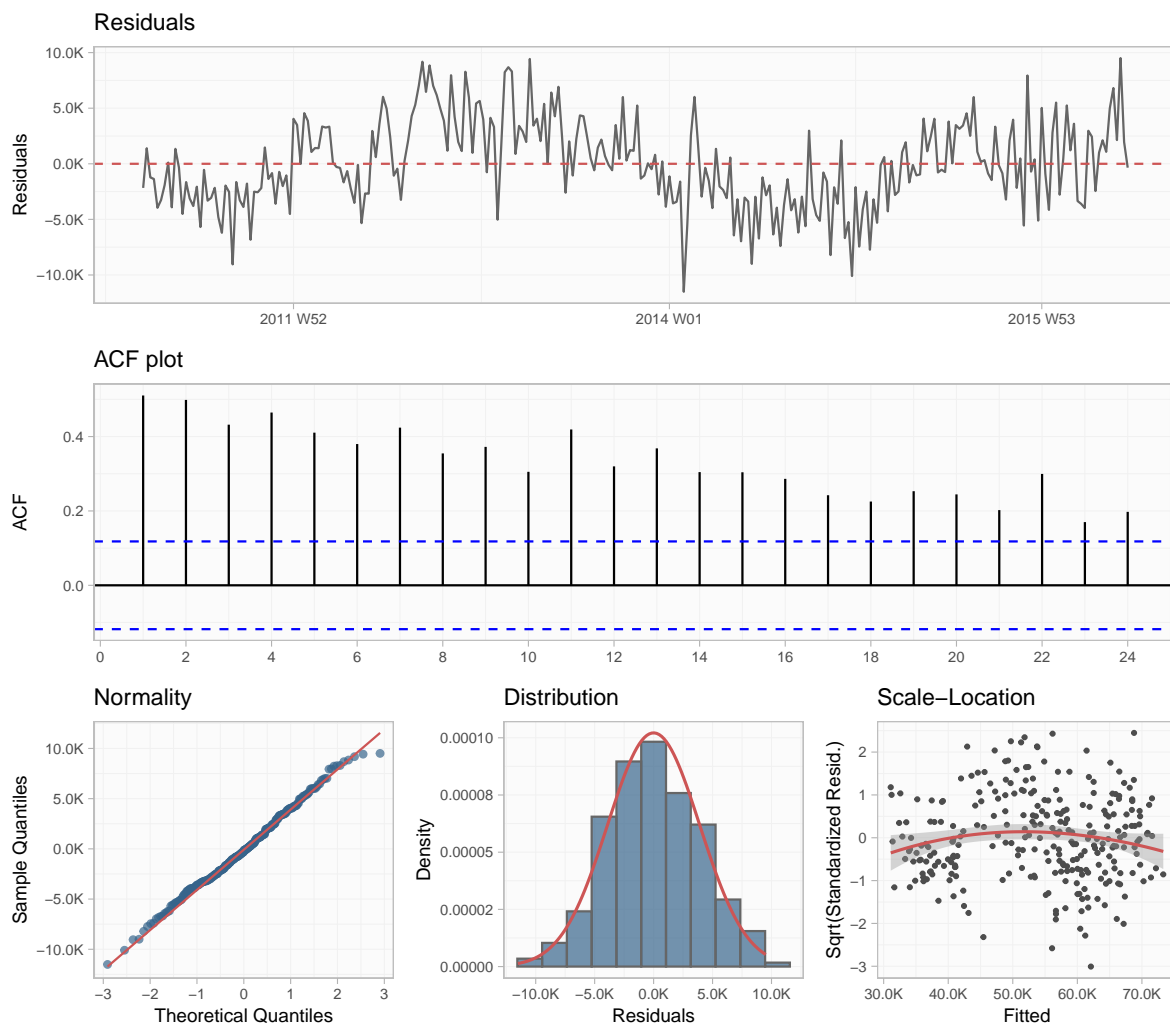


Figure 6.7.: Residuals diagnostic plots of the initial HOUSEHOLD regression model fit

To add to the above observations, we get inconclusive results for the stationarity of the model from the ADF and KPSS tests (Table 6.7), while clearly the Ljung-Box confirms the presence of autocorrelation in residuals. The normality and heteroscedasticity tests (Shapiro-Wilk, Breusch-Pagan) cannot be trusted at this stage, since our model does not meet most of the regression essential assumptions.

Test	Statistic	P-value
KPSS	0.399	0.0776
ADF	-2.544	0.347
Ljung-Box	72.707	< 0.01
Breusch-Pagan	0.695	0.4045
Shapiro-Wilk	0.995	0.5157

Table 6.7.: Tests results on the residuals of the initial HOUSEHOLD model fit

Overall the initial model has significant deficiencies in explaining the temporal structure, variance and distribution of residuals. To fix the issues and create a robust regression, we must refine the model in the following section.

6.4.3. Refined Regression Model

To address the deficiencies of the earlier HOUSEHOLD model, we will add three lagged sales predictors (lag 1 - lag 3) as specified in previous chapters. Thus the new model will be:

$$\begin{aligned}
 \text{Sales}_{\text{HSHLD},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} + \\
 & \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Event} + \beta_{16} \cdot \text{Price} + \beta_{17} \cdot \text{Lag 1} + \\
 & \beta_{18} \cdot \text{Sales Lag 2} + \beta_{19} \cdot \text{Sales Lag 3} + \epsilon_t
 \end{aligned} \tag{6.5}$$

where:

- Lag 1 is the sales value one period before t
- Lag 2 is the sales value two periods before t
- Lag 3 is the sales value three periods before t

6.4.4. Residual Diagnostics for the Refined Model

This time residuals fluctuate more randomly around zero with no apparent systematic trends or other cyclical behavior. The amplitude of residuals appears more stable over time, indicating that variance is handled better after the introduction of the lagged predictors. This insight adds

to the evidence that there is strong short memory in the dataset. ACF plot shows a clear improvement over the initial model, with the absence of patterns and only random peaks, indicating an adequate capture of the temporal dependencies through the use of lags.

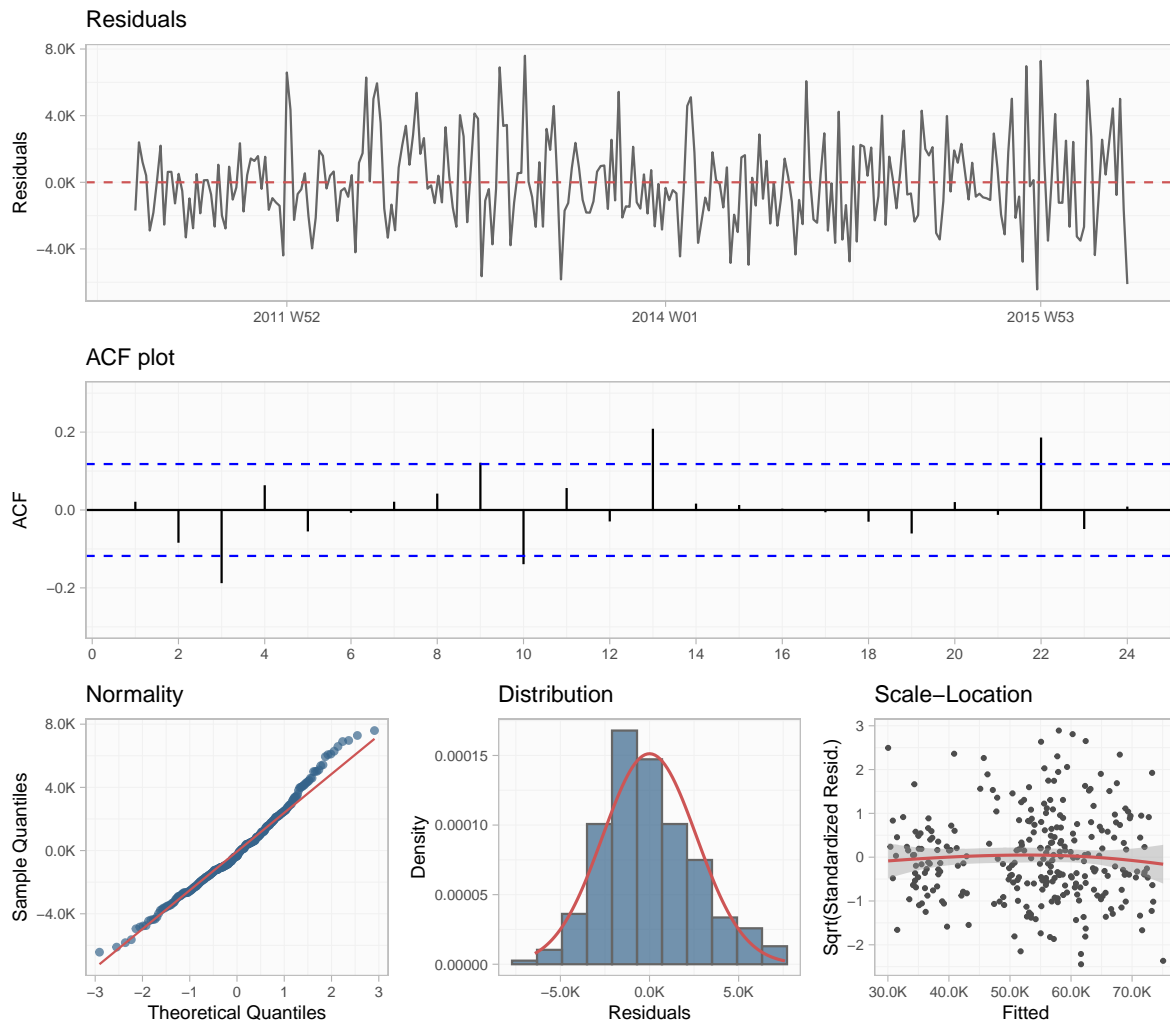


Figure 6.8.: Residuals diagnostic plots of the refined HOUSEHOLD regression model fit

Moving on, the assumption of normality seems to be met, with a roughly bell-shaped distribution in the histogram and a only some minor issues in the Q-Q plot at the upper end of the diagonal that are unlikely to significantly affect our model's validity. Last, the scale-location plot shows a better fit this time, indicating that we have constant variance across the fitted values, supporting the assumption of homoscedasticity of residuals. In addition, the combination of ADF and KPSS tests in Table 6.8 confirm the stationary behavior of our residuals while the result from the Ljung-Box test (p-value 0.7231) indicates that the refined model has indeed addressed the temporal dependencies in the data. Breusch-Pagan test is slightly above the 0.05 significance level and

combined with the visual inspection from the scale-location plot, we can conclude that residuals are homoscedastic.

Test	Statistic	P-value
KPSS	0.129	0.1
ADF	-6.709	< 0.01
Ljung-Box	0.126	0.7231
Breusch-Pagan	3.194	0.0739
Shapiro-Wilk	0.986	< 0.01

Table 6.8.: Tests results on the residuals of the refined HOUSEHOLD model fit

The p-value result from Shapiro-Wilk test (< 0.01) leads us to reject the null hypothesis of normality in residuals, however based on the QQ-plot, it is evident that the Shapiro-Wilk is detecting the minor deviations at the upper tail. In this case, the non-normality is not severe and the model's results remain valid, even if the residuals are not perfectly normal.

6.4.5. Fit Assessment

According to the model fit summary in Table 6.9, the model explains an impressive 95% of the variation in sales ($R^2 = 0.95$, adjusted $R^2 = 0.947$), while the overall p-value indicates that we have a statistically significant model. The Trend variable is found significant, suggesting that sales for HOUSEHOLD products increase on average by about 17 units. GVIF value is found relatively high but within the acceptable limits. This inflated value is caused by the introduction of the lagged sales predictors, which tend to be highly correlated with the trend variable, as both capture time-related patterns.

Moving on to the Month dummies we can infer that annual seasonality affects sales, while the end-week dummy, indicates that as month progresses, sales drop, confirming our initial assumptions regarding the presence of a secondary cycle within each month. SNAP days also tend to increase sales (+1,702 units) probably as a side effect of the increased visits during these days. Again we see that Events are not found statistically significant and do not have any impact on HOUSEHOLD sales, as seems to be the case for the Price variable too. Last, all the newly introduced lags (Lags 1 to 3) are found to be statistically significant, indicating strong persistence in sales. This outcome supports the decision to include lagged predictors, as the short-term memory of this series appears to play an important role.

Refined HOUSEHOLD Model				
Variable	Estimate	95% CI	P-value	GVIF
(Intercept)	24,892	-1,891, 51,676	0.068	
Trend	17	4.7, 30	0.007	9.6
Month dummies				3.3
<i>Jan</i>	—	—		
<i>Feb</i>	3,901	2,055, 5,746	<0.001	
<i>Mar</i>	-898	-2,875, 1,080	0.372	
<i>Apr</i>	-674	-2,585, 1,236	0.487	
<i>May</i>	-1,119	-2,935, 696	0.226	
<i>Jun</i>	793	-954, 2,540	0.372	
<i>Jul</i>	79	-1,802, 1,961	0.934	
<i>Aug</i>	1,382	-542, 3,305	0.159	
<i>Sept</i>	-334	-2,304, 1,635	0.738	
<i>Oct</i>	-1,722	-3,692, 249	0.087	
<i>Nov</i>	-2,445	-4,264, -626	0.009	
<i>Dec</i>	-2,311	-3,999, -624	0.007	
End-Week dummy				2.1
0	—	—		
1	-1,285	-2,219, -352	0.007	
SNAP dummy				2.1
0	—	—		
1	1,702	759, 2,645	<0.001	
Event dummy				1.2
0	—	—		
1	-59	-771, 653	0.871	
Price	-4,016	-9,257, 1,226	0.133	2.0
Lag 1	0.16	0.04, 0.27	0.009	18
Lag 2	0.39	0.28, 0.50	<0.001	17
Lag 3	0.31	0.19, 0.42	<0.001	18
<i>R</i> ²	0.950			
Adjusted <i>R</i> ²	0.947			
Statistic	257			
<i>df</i>	19			
<i>p</i> -value	<0.001			

Abbreviations: CI = Confidence Interval, GVIF = Generalized Variance Inflation Factor

Table 6.9.: Summary of the refined regression model for the HOUSEHOLD time series

Overall, we have a robust model, that meets the fundamental regression assumptions and on top, demonstrates strong performance by successfully capturing the key factors that drive the demand for the products in the HOUSEHOLD category.

To conclude, we see in Figure 6.9, that the fitted line has a close fit on the actual sales, doing a good job in capturing the trend and seasonality. Despite some extreme fluctuations, where the model as it is defined, cannot fully capture, we overall have a solid performance without systemic biases giving the confidence that the underlying model is well-suited for explaining the sales dynamics in the HOUSEHOLD category.

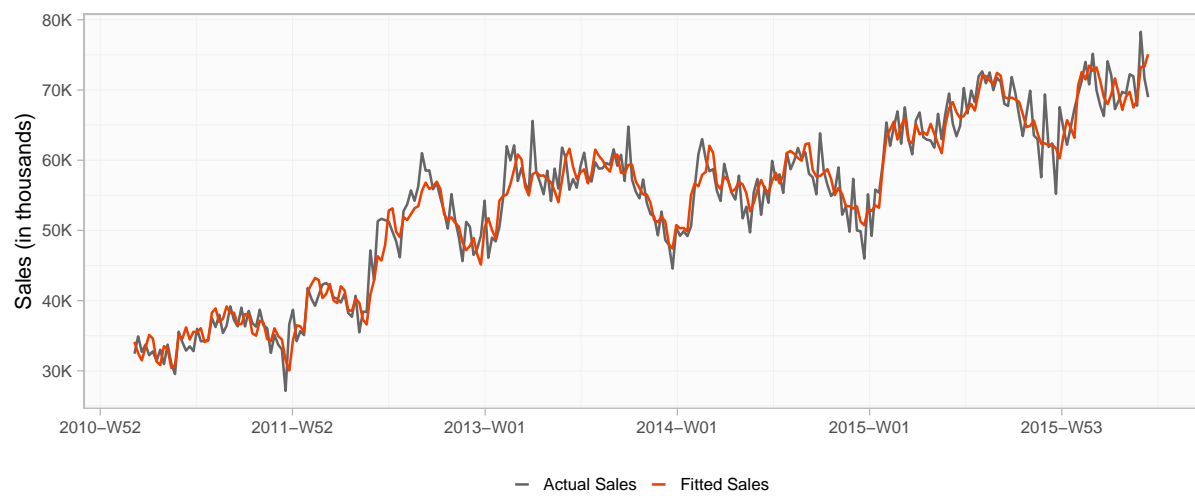


Figure 6.9.: Plot of the refined HOUSEHOLD model fit vs original values

7. Out-of-Sample Model Evaluation

7.1. Dataset Preprocessing

In this final stage of the project, following the application of regression models in the previous chapter, we will evaluate their performance by comparing them with alternative time series modeling techniques. The goal is to assess their predictive capabilities in terms of their out-of-sample forecasting accuracy.

Before we enter the modeling phase, we need to go through an extra data processing step. The three datasets will be split into two periods. The first period, also referred to as the training set, will be used to fit and validate the models. The second period, called the test set, will be used to assess the out-of-sample performance of the models. We will follow the commonly used 80-20 rule, meaning that 80% of the data will be allocated to the training set and the remaining 20% will become the test set. This ratio makes sure that there is enough historical data for learning while keeping enough data for evaluation.



Figure 7.1.: Training - Test split process

The training set will allow the models to learn how to capture underlying patterns and relationships within the time series while the test dataset is used to ensure unbiased results and avoid overfitting. Overfitting occurs when a model learns the noise from the training set rather than the real time series structure making it difficult to generalize on new unseen data. The purpose of the test set is to represent the future and thus provide an independent evaluation of the models.

The way we divide datasets into training and test sets is similar to how forecasting is practiced in the real world, where we must predict future values using only past observations.

Following the preprocessing performed prior to the regression analysis, our datasets comprise of 276 weeks from 2011 W10 to 2016 W24. Applying the 80-20 split, the training set will have 221 weeks of data, covering the period 2011 W10 to 2015 W27. The remaining 55 weeks, from 2015 W28 to 2016 W24, will become the test set for the out-of-sample evaluation. This temporal split ensures that the evaluation respects the chronological order of the data, preventing future values from leaking into the training process. This arrangement not only improves the reliability of the evaluation but also provides a realistic measure of a model's ability for generalization.

7.2. Modeling Process

With the training and testing sets for each time series prepared, we now transition to the modeling phase. The main goal here is to develop predictive models that can effectively predict the values in the test set. These models will be evaluated on their predictive performance but also in comparison to one another so we can understand their relative strengths and weaknesses. Each predictive model offers different advantages and this analysis will try to assess their performance while answering on how predictable are the time series under investigation.

To achieve this, we will begin with Simple Exponential Smoothing (SES) model, which will serve as a baseline benchmark. This model, while straightforward, provides an important reference point for evaluating the added value of the other two approaches. Next, we will implement Holt-Winters model, which explicitly accounts for the observed seasonal and trending behavior in the data, offering a more nuanced approach to time series forecasting. Finally, we will apply the regression models developed in Chapter 6. In this step, we will further refine these models by excluding any non-significant variables specific to each dataset, ensuring their performance and compliance with the OLS method's assumptions.

All models will be fit on the training set and then create predictions on the test set, where we will assess their performance. For each time series we will plot the results over the test sets actual values and visually assess them along with their residuals. Finally, we will quantify and compare the performance of all time series models out-of-sample predictions using the MAE, MAPE and RMSE accuracy measures.

7.2.1. FOODS Models

Starting from the FOODS time series the SES will be trained using the equation

$$\hat{\text{Sales}}_{\text{FOODS},t} = \alpha \cdot \text{Sales}_t + (1 - \alpha) \cdot \hat{\text{Sales}}_{t-1} \quad (7.1)$$

where the calculated $\alpha = 0.2032$.

Following the same process, the Holt-Winters model will be trained using the equation:

$$\hat{\text{Sales}}_{t+h} = (l_t + h \cdot b_t) \cdot s_{t+h-p(k)} \quad (7.2)$$

with calculated parameters $\alpha = 0.44$, $\beta = 0.001$, $\gamma = 0.92$.

Last, for training the regression model we will use the equation:

$$\begin{aligned} \text{Sales}_{\text{FOODS},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} + \\ & \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Lag } 1 + \beta_{16} \cdot \text{Lag } 4 + \varepsilon_t \end{aligned} \quad (7.3)$$

The results of the models are presented in Figure 7.2, which displays both the in-sample period (training set) and the out-of-sample period (test set) along with the resulting predictions for each model. A visual inspection reveals that the Simple Exponential Smoothing (SES) model performs well in capturing only the mean of the recent weeks, but it fails to capture any cyclical and trending behaviors observed in the test set. In contrast, the Holt-Winters model (HW) effectively captures the annual and monthly seasonality as well as the trend, but it fails to adapt to the changing level that starts in the first weeks of 2016. This limitation is expected, as the Holt-Winters model is designed to extrapolate past patterns into the future and requires a buffer period when new patterns are introduced. As for the Linear Model (LM), it demonstrates a strong fit on the test set and although it slightly underestimates the magnitude of some peaks and troughs, it is the only model that successfully adjusts to both seasonal patterns and the trend shifts observed starting in 2016. If compared to the Holt-Winters model, where it follows a steeper drop for the rest of 2015 and struggles to keep up with the change of trend, the LM is better at generalizing its predictions for both the second part of 2015 and the new patterns of 2016.

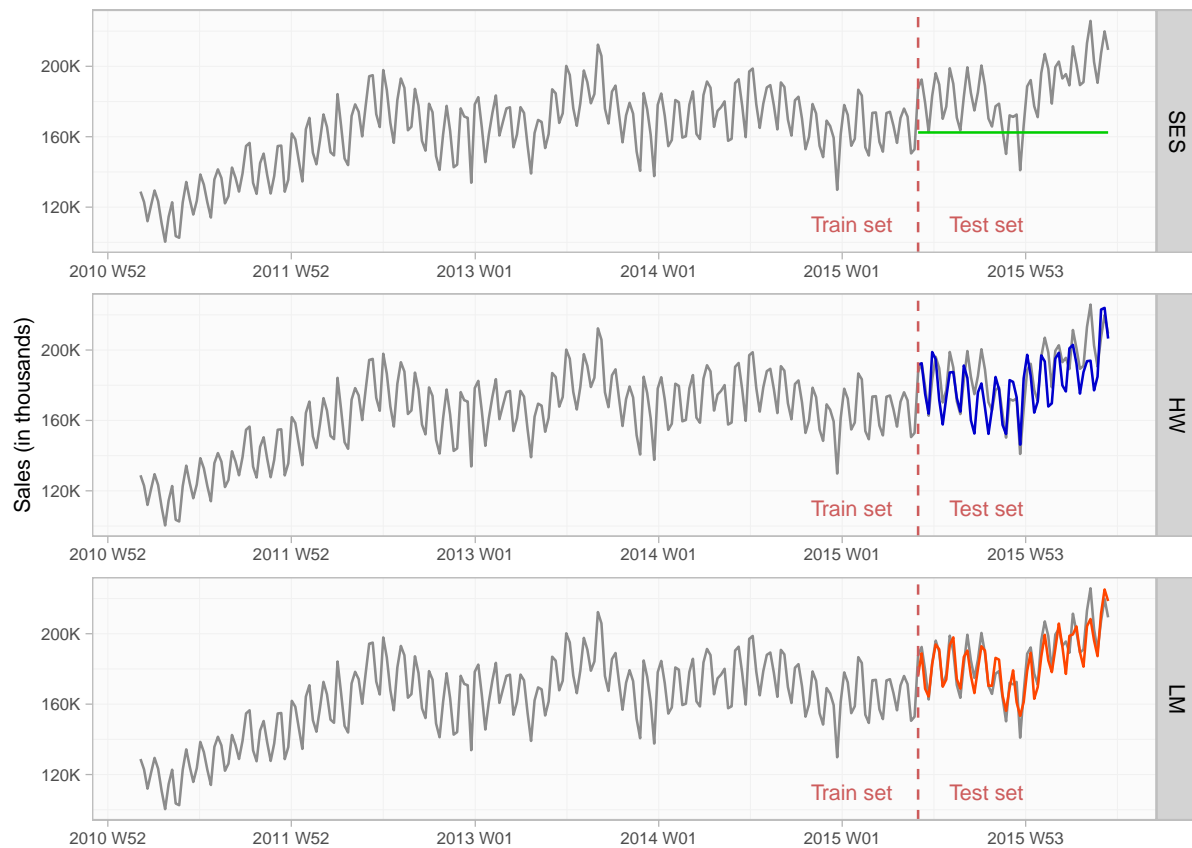


Figure 7.2.: Plot of the FOODS models fit on the test set

The residual plots in Figure 7.3 confirm that the SES model tends to under-forecast the test set values as almost all of the points are located above the zero line on the upper half of the plot. For this model, there is a noticeable trend as well as seasonal behavior that can be observed in the residuals leading us to conclude that results are biased. Moving to the Holt-Winters model, we see a more stationary behavior in the residuals, with a tendency to under-forecast after the beginning of 2016. In comparison to the residuals from SES, it shows more randomness as the model has successfully captured to some extent the annual patterns. Regarding the Linear Model we have an obvious better overall performance. Notably, we can see that the magnitude of the residuals is smaller and more tightly clustered around the zero mean line. Residuals appear to be randomly scattered around zero with no systematic error behavior. Again, if we focus on the period after the beginning of 2016 this model as well struggles to fit the abrupt change of trend and is slightly under-forecasting but not to the same extent as the two preceding models. Overall it is a better fit demonstrating stationary behavior and smaller error values.

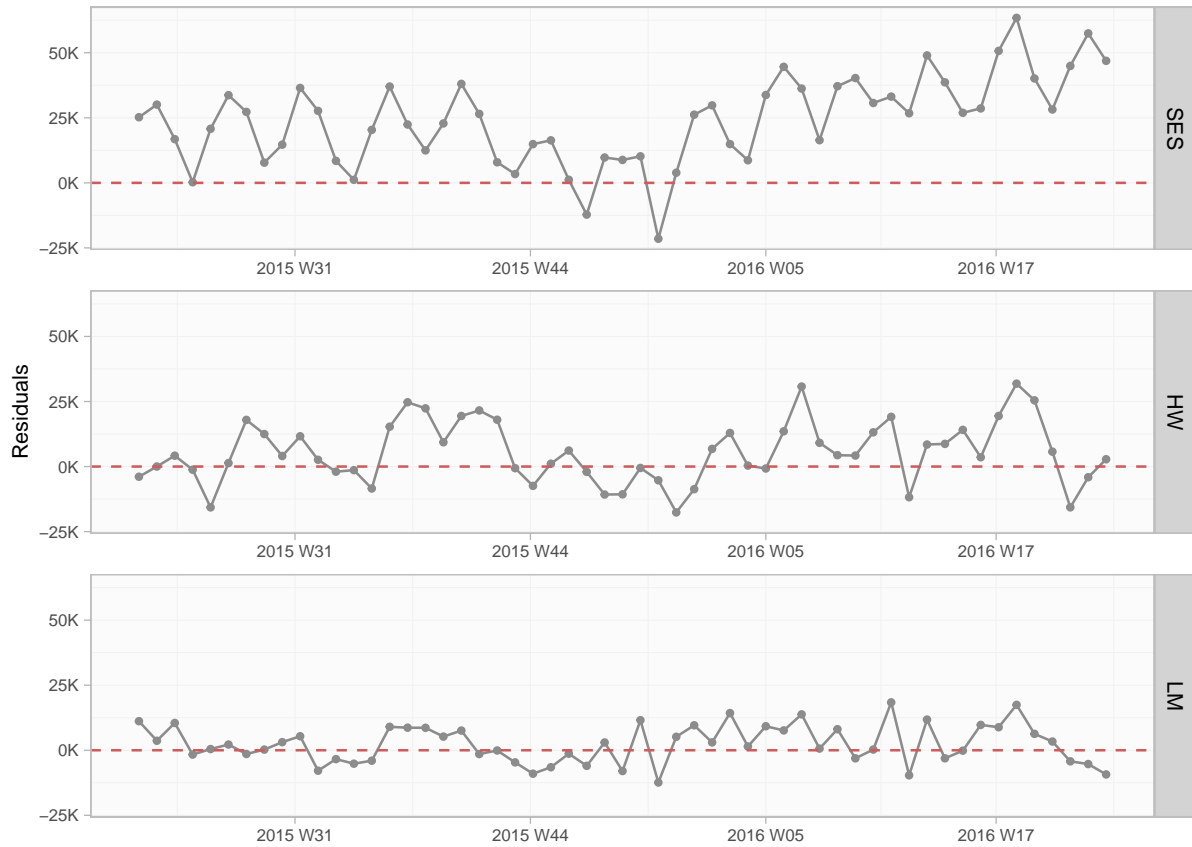


Figure 7.3.: Plot of the FOODS models residuals from the test set

We can conclude from the visual inspection that for the FOODS time series the most appropriate model is the Linear Model, by succeeding on generalizing the in-sample patterns into the future, adapting well to changes and proving that the FOODS time series can produce reliable predictions. We will return to quantify the predictability of the time series after concluding the modeling of the other two datasets, now moving on to the HOBBIES time series.

7.2.2. HOBBIES Models

For the HOBBIES time series the SES model will be trained with the equation:

$$\hat{\text{Sales}}_{\text{HOBBIES},t} = \alpha \cdot \text{Sales}_t + (1 - \alpha) \cdot \hat{\text{Sales}}_{t-1} \quad (7.4)$$

with an $\alpha = 0.3694$.

Following, the Holt-Winters model will be applied to the training data with the equation:

$$\hat{\text{Sales}}_{\text{HOBBIES},t+h} = (l_t + h \cdot b_t) \cdot s_{t+h-p(k)} \quad (7.5)$$

with parameters $\alpha = 0.28$, $\beta = 0.01$, $\gamma = 0.77$.

Third, the regression model that will be used is:

$$\begin{aligned} \text{Sales}_{\text{HOBBIES},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{Price} + \\ & \beta_{17} \cdot \text{Lag } 1 + \beta_{14} \cdot \text{Lag } 2 + \beta_{15} \cdot \text{Lag } 3 + \varepsilon_t \end{aligned} \quad (7.6)$$

From the visual inspection of the fit shown in Figure 7.4, the SES model appears to perform poorly in predicting the sales during the out-of-sample period. The prediction line remains flat and does not follow the annual patterns of the actual sales. This suggests that SES is not capturing the trend well and is instead only forecasting the mean of the historical data, missing the dynamic behavior particularly during the last few weeks of 2015 and 2016, where sales rise. Compared to the FOODS time series, HOBBIES sales have a difficult pattern with changes of level and trend throughout both the training and the test sets, as well as shocks that are not easy to explain. This makes the SES modeling rather challenging but as a benchmark model, does a good job at predicting at least the level of the out-of-sample period.

The Holt-Winters model does a better job in capturing trend and seasonality to some extent, as its prediction line follows closer the out-of-sample values. However, it still under-performs in tracking the sharp changes within each year, as the prediction line misses some of the upward and downward movement in the actual sales. Although this model is better at handling seasonality, the seasonal component of the HOBBIES time series is rather weak and not stable throughout the dataset, making it difficult to extrapolate it to the out-of-sample period. If we combine the above with the multiple changes in level and the shocks that are present throughout the data, the model struggles to predict the test sets sales. Moving to the Linear Model, the results show a much closer fit to the actual test set sales. While there is some discrepancy, particularly in the large peaks and troughs that occur, it adapts better compared to the other models. However, it still misses some of the magnitude of the spikes, suggesting that the variables used in the regression are not able to fully explain these shocks.

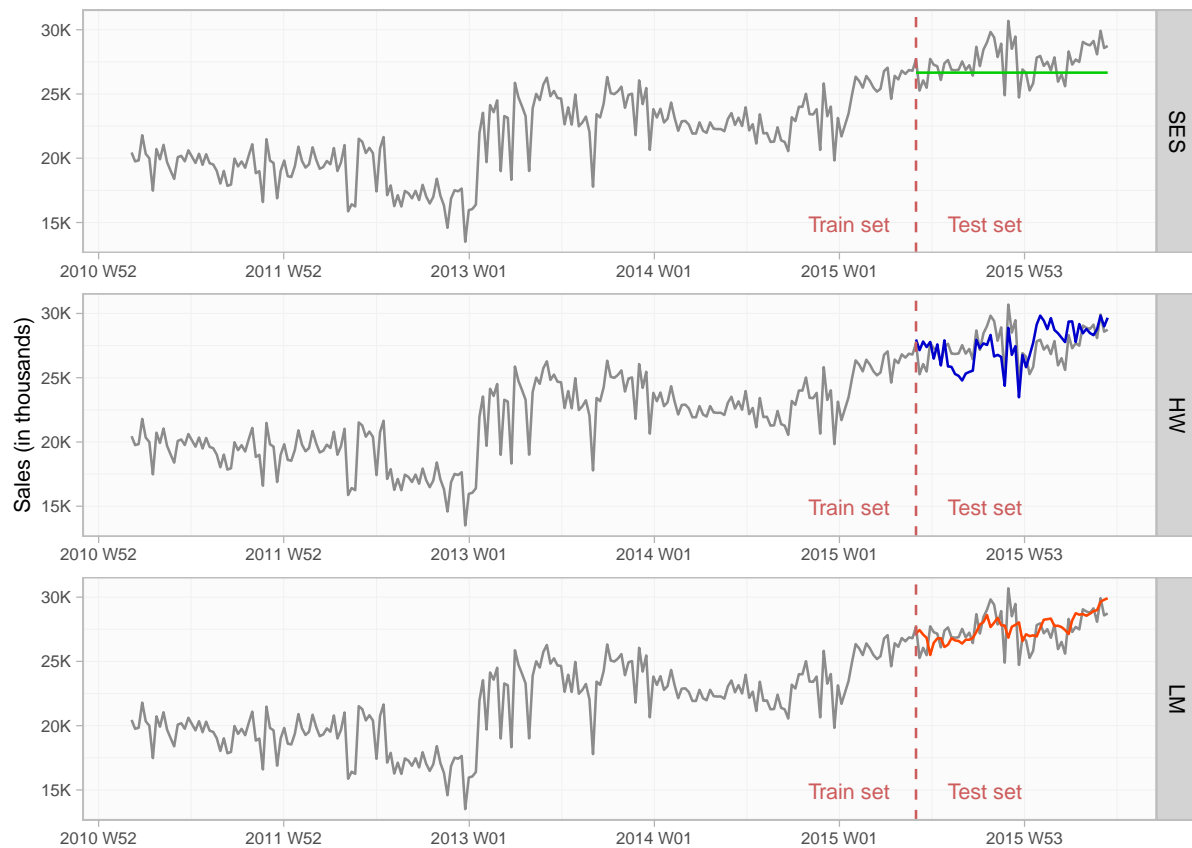


Figure 7.4.: Plot of the HOBBIES models fit on the test set

The residual plots for the three models show different behaviors. For the SES model, the residuals exhibit a cyclical pattern indicating that the model is not fully capturing the seasonality and trend in the data. The big shocks at the end of 2015 are found difficult to be captured from a simple exponential model and overall, the residuals fall on the upper half of the zero line indicating that the model is under-forecasting (bias) the test sets sales. The Holt-Winters model shows similar behavior but this time the cyclical pattern is more distinguishable. These patterns imply that the model under-forecasts during some periods and over-forecasts during others. The noticeable fluctuations in residuals, lead us to accept that the model fails in accurately fitting both trend and seasonality of the out-of-sample period and despite fewer large patterns compared to SES, the residuals remain non-random. The LM, on the other hand, demonstrates the best performance among the three models. The residuals of this model are more sporadically distributed around zero with less large patterns, therefore, a better fit. While some larger deviations remain, these are less, implying that the LM adapts better to the underlying data patterns.

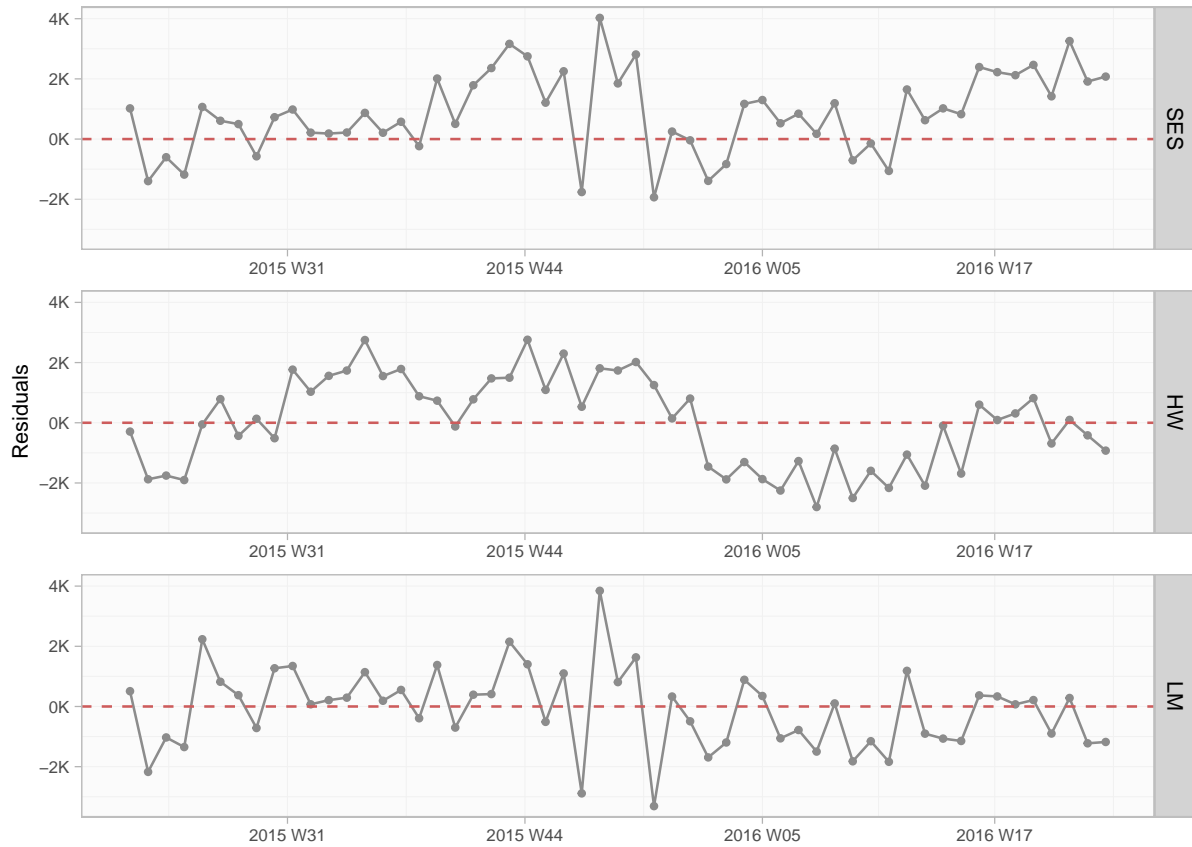


Figure 7.5.: Plot of the HOBBIES models residuals from the test set

We can conclude from the visual inspection that both Exponential models show significant systematic errors, indicative of poor fitting. Specifically the performance of the Holt-Winters model is not much improved compared to the benchmark model as it struggles with trend and seasonality shifts. On the other side, the Linear Model provides the best fit out of the three, though occasional larger deviations suggest there is still some unexplained behavior in the sales of the HOBBIES category.

7.2.3. HOUSEHOLD Models

For the HOUSEHOLD time series the SES model will be trained with the equation:

$$\hat{\text{Sales}}_{\text{HSHLD},t} = \alpha \cdot \text{Sales}_t + (1 - \alpha) \cdot \hat{\text{Sales}}_{t-1} \quad (7.7)$$

with an $\alpha = 0.5667$.

Following, the Holt-Winters model will be applied to the training data with the equation:

$$\hat{\text{Sales}}_{\text{HSHLD},t+h} = (l_t + h \cdot b_t) \cdot s_{t+h-p(k)} \quad (7.8)$$

with parameters $\alpha = 0.11$, $\beta = 0.01$, $\gamma = 0.99$.

Third, the regression model that will be used is:

$$\begin{aligned} \text{Sales}_{\text{HSHLD},t} = & \beta_0 + \beta_1 \cdot \text{Trend}_t + \sum_{m=2}^{12} \beta_m \cdot \text{Month}_m + \beta_{13} \cdot \text{End-Week} + \\ & \beta_{14} \cdot \text{SNAP} + \beta_{15} \cdot \text{Lag 1} + \beta_{16} \cdot \text{Lag 2} + \beta_{17} \cdot \text{Lag 3} + \varepsilon_t \end{aligned} \quad (7.9)$$

Results of the out-of-sample performance in Figure 7.6 show a better fit than the ones of the HOBBIES category. The HOUSEHOLD time series demonstrates a more stable behavior with clear patterns and trend throughout the train set. These characteristics of the series help our

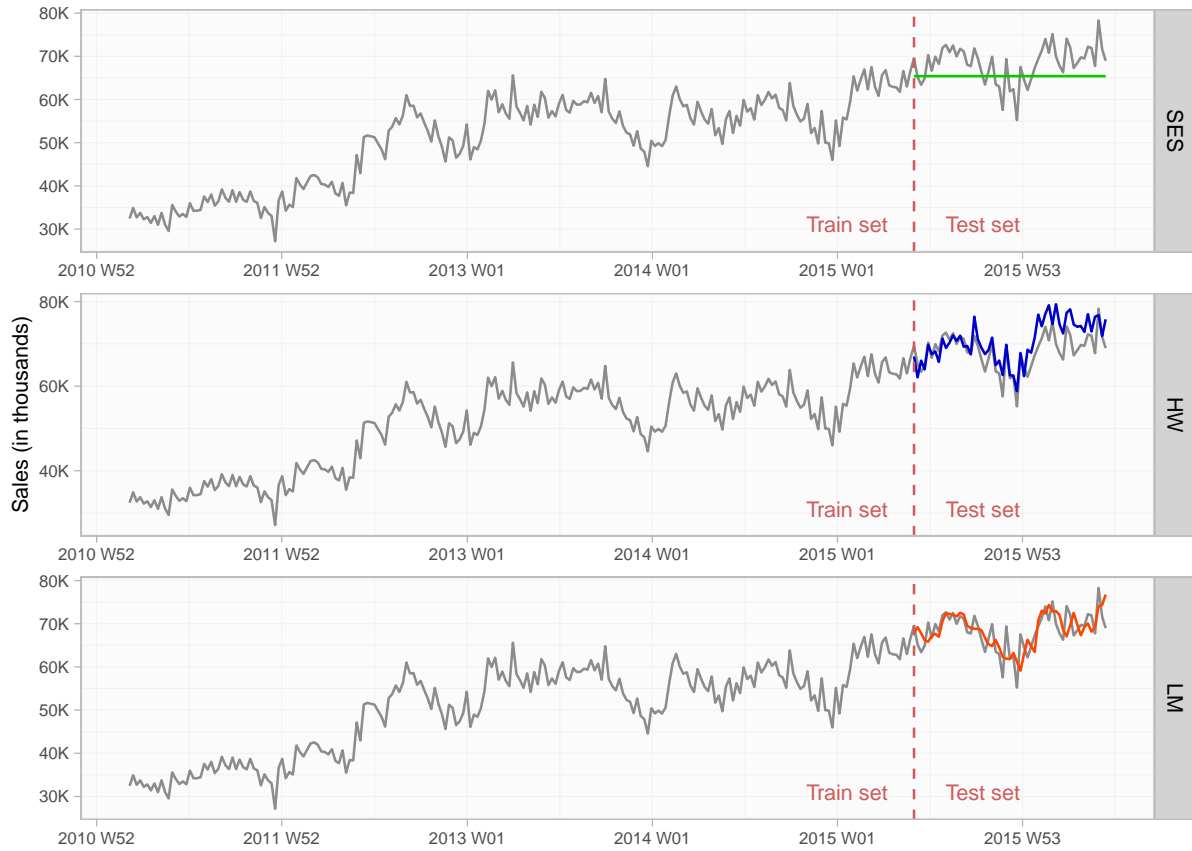


Figure 7.6.: Plot of the HOUSEHOLD models fit on the test set

benchmark model perform well on predicting the level of the test period. Moving to the Holt-Winters, the model appears to capture the general trend as the prediction line follows a similar upward movement as the test data. Seasonal patterns seem to be adequately modeled, as the fluctuations in the predictions align with the variations in the actual data. Again we see some un-modeled shocks that are not part of the historical pattern, thus the predictions demonstrate some deviations in amplitude for some observations, especially during 2016. Compared to both previous models, the Linear Model has a tight fit along the actual sales, succeeding in keeping track of the cyclical and trending behavior. Even though there is an issue with the shocks observed in the end of 2015, where it fails to capture some irregular spikes and dips, it demonstrates a closer fit for year 2016 compared to the Holt-Winters model.

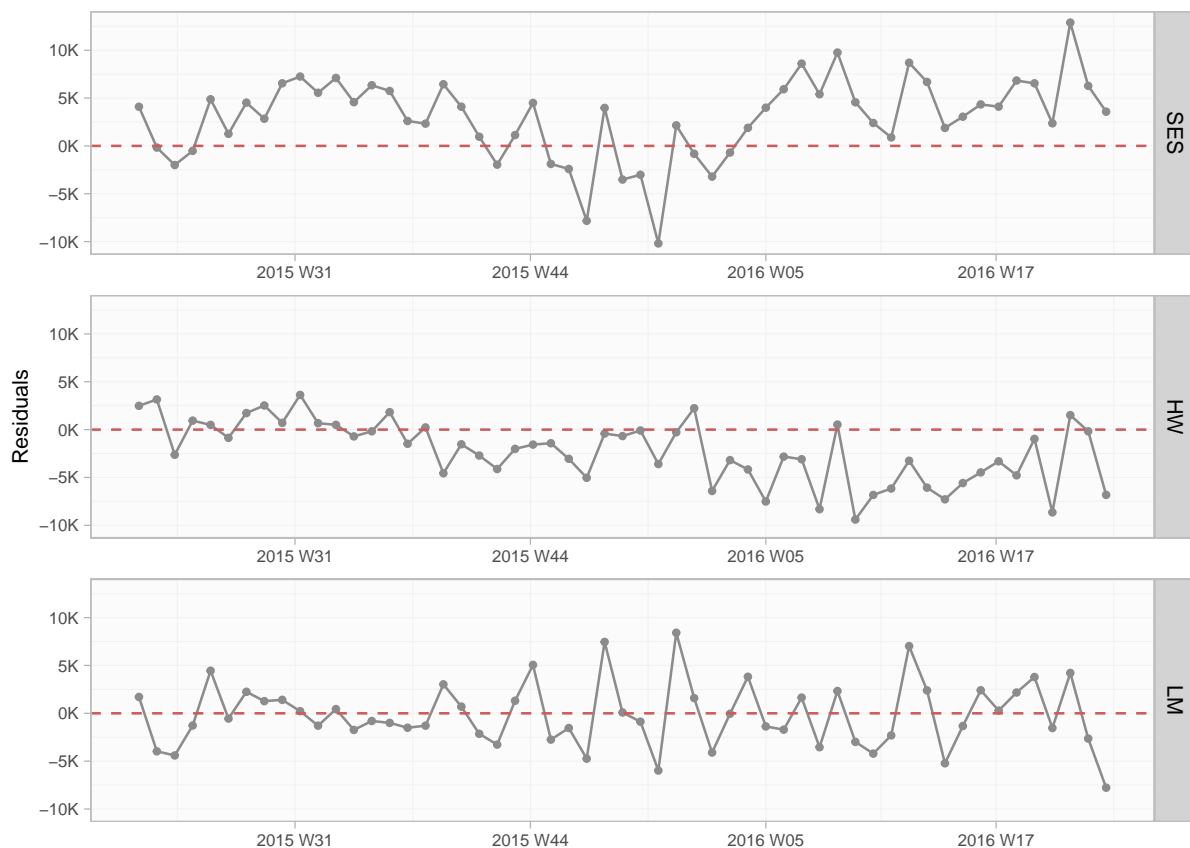


Figure 7.7.: Plot of the HOUSEHOLD models residuals from the test set

The above can be confirmed from the residual plots where the benchmark model demonstrates signs of cyclical patterns and is mostly under-forecasting the test set. Interestingly, the Holt-Winters has a good start for the most part of 2015, but then follows a downward movement in the residuals, leading the model to over-forecast while the residual's amplitude slightly increase. Finally, the Linear Model's residuals are centered around zero showing signs of stationarity,

suggesting that overall the model created unbiased predictions. Similar to the HW model, there is an increase in amplitude emerging after 2016, due to the appearance of new patterns that were not present in the training set. Overall we can conclude from the visual inspection of the residuals plots that the Holt-Winters seems to have a slightly better fit than the benchmark Simple Exponential model but Linear Model residuals show stationary behavior with no significant patterns, making it once more the best fit among the three.

After concluding the visual inspection of the out-of-sample results of all three categories models, we will proceed to the final part of the evaluation, where we will quantitatively measure the models predictions accuracy and assess their overall performance.

7.3. Accuracy Evaluation

In this last section, we move to the assessment of the models' predictive performance, to find out what is the typical accuracy rate at which weekly sales can be forecasted. While the visual inspection of out-of-sample results provided significant insights on the models performance, this part aims to offer a more objective and measurable comparison. Accuracy metrics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), will be employed on the testing periods' data, to evaluate how well each model captures the underlying patterns and variability. By quantifying the magnitude and nature of prediction errors, we will get more details about each model's strengths, limitations and overall suitability for the given dataset.

MODEL	FOODS			HOBBIES			HOUSEHOLD		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
SES	25,328	13.04%	29,365	1,294	4.61%	1,582	4,322	6.25%	5,103
HW	10,094	5.32%	12,918	1,249	4.57%	1,466	3,079	4.54%	3,952
LM	6,313	3.37%	7,696	1,041	3.82%	1,307	2,680	3.96%	3,343

Table 7.1.: Accuracy results for the in-sample fit of the models

Starting from the FOODS category, the SES model has the largest errors across all metrics, proving our earlier visual assessment of facing struggles to capture variability and any kind of patterns in the time series. For the same category, the HW model outperforms SES, vastly reducing the errors, as the ability of this model to capture trend and seasonality is significant for this category. The LM manages to get even better results, outperforming both Exponential Smoothing models. The LM achieves the lowest errors across all metrics, making it the most

effective model out of the three for predicting FOODS sales. Moving to the HOBBIES category, once again the SES model has the weaker results but in this case, results are better compared to the SES from FOODS category. The HW model is slightly better than the SES, with a minor reduction in errors, however the improvement is so small that we cannot say with confidence that it is a better model than the SES. The LM again is performing better than both the other models, suggesting a stronger adaptability to the difficult case of the HOBBIES time series. Last, we see the same pattern now in the HOUSEHOLD category, where the SES has the highest errors, followed by the HW that demonstrates a noticeable improvement over SES. Finally once more, the LM provides the best performance among the three models.

Overall we see that SES, as expected from a benchmark model, creates some decent but not adequate predictions for all three time series. The HW model, appears to be significantly better than SES, especially in the FOODS category, where it leverages its ability to account for trend and seasonality. However, its performance is slightly less competitive in the HOBBIES and HOUSEHOLD categories. The most accurate model is the LM, achieving a 3.37% MAPE for FOODS, 3.82% for HOBBIES and 3.96% for HOUSEHOLD, meaning its forecasts are on average over 96% accurate across categories. This demonstrates a high level of accuracy in the predictions for weekly sales, suggesting that trend and seasonality together with the use of explanatory variables for prices and promotions, effectively capture the dynamics of the product category time series.

8. Main Findings and Conclusion

In this final part we bring together the outcomes and results of this study to answer the main research questions raised in the beginning and provide a summary of key conclusions.

8.1. Research Questions and Findings

What are the demand patterns across product categories?

The analysis revealed distinct demand patterns across FOODS, HOBBIES and HOUSEHOLD categories. The FOODS category showed strong upward trend with prominent annual seasonality. A second week-of-month cycle was also observed, reflecting consumer budgeting behavior. HOUSEHOLD demonstrated a gradual upward trend but with moderate seasonal fluctuations, indication of stable demand. HOBBIES, meanwhile, demonstrated irregular patterns with level shifts and considerable volatility, driven by one-off factors and external shocks. These findings highlighted the necessity for category-specific modeling strategies.

What are the critical factors influencing retail sales behavior?

The regression analysis identified key factors such as pricing, SNAP promotions and temporal cycles as important. Contrary to our initial assumptions, the price variable was the one that was found to have minimal impact on FOODS and HOUSEHOLD compared to the slightly higher sensitivity observed in HOBBIES. SNAP promotions significantly affect FOODS and HOUSEHOLD sales, while calendar events and holidays had minimal impact across categories. Temporal factors like trends and seasonality were found to be the most impactful to all categories, offering actionable insights for creating predictions.

How forecastable are retail sales time series?

Forecastability varies across categories. FOODS regression showed the best out-of-sample predictability, with a MAPE result of 3.37%, mainly due to its stable temporal patterns. HOUSEHOLD and HOBBIES regression models achieved a MAPE of 3.82% and 3.96% respectively, creating overall impressive prediction results. The forecast accuracy for all series, significantly

improved with the inclusion of lagged variables, especially HOBBIES, which confirms the importance of short-term dependencies in dynamic categories.

How do deterministic and stochastic components interact in retail sales time series?

The decomposition and regression analysis highlighted the complex interactions between deterministic components (trends and seasonality) and stochastic elements (residuals). In FOODS and HOUSEHOLD, deterministic structures explained most of the variance, while stochastic components played a less prominent role. HOBBIES, however, showed increased levels of noise, leading us to the assumption that other external factors influence these series. The integration of lagged variables successfully captured the interactions between structured patterns and short-term fluctuations, enhancing the overall model performance.

8.2. Conclusion

This study tried to prove that detailed exploratory analysis and careful interpretation of time series characteristics are essential for generating reliable demand forecasts. The findings underline the importance of recognizing category-specific behaviors and deterministic and stochastic components to explain the behavior of time series data. While FOODS and HOUSEHOLD categories exhibited predictable structures, HOBBIES category is found more difficult to be modeled with the current set of variables.

The regression models applied in this study and with the addition of lagged variables achieved high predictive accuracy, highlighting the value of methodical preparation and analysis. The conclusions provide a better understanding of what could possibly drive demand for retail and the methods used throughout our research could offer practical examples for demand planning personnel in regards to making informed decisions for inventory management, promotional strategies and supply chain optimization.

Regarding the product categories we studied, FOODS stands out as the most predictable with a stable progressive growth, strong seasonal behavior and socio-economic cycles (SNAP) that was successfully identified as significant in modeling. Demand for HOUSEHOLD products is consistent and prices or promotions have only marginal effects. Hobbies however were found more complex. Their demand is not as stable, as it can be impacted by some supply interruptions and also more price-sensitive; thus, their forecasting can be further improved with additional information.

The limitations of the study are mainly attributed to methodological choice. The weekly aggregation choice served the exploration of the general drivers, but it could mask daily patterns or other shorter-term fluctuations in demand. Moreover, the use of merely statistical methods although found capable in identifying the general trend and seasonality, may not be the best to capture a range of complex non-linear relationships that are commonly present in retail data. Using machine learning methods could help learn complex relationships and patterns that are not explained by the current statistical approach. In the case of the Events variable, the weekly aggregation amplified the effect on sales, where an analysis at the daily level could be proved more significant. Finally, the use of weekly mean price as a representative measure may fail to capture price volatility and promotional effects, potentially masking critical pricing dynamics that influence consumer behavior.

8.3. Recommendations for Future Research

- **Incorporating External Variables:** Integrate macroeconomic indicators, weather data and consumer sentiment indices to enhance demand modeling, especially for dynamic categories like HOBBIES.
- **Adopting Machine Learning Techniques:** Explore advanced methods such as neural network techniques (RNN, LSTM) and ensemble models to capture non-linear relationships and complex interactions. These flexible models, though, require careful optimization of their architecture (e.g., number of layers or neurons) to avoid overfitting. This could be achieved with the use of automated selection methods (see Thomaidis & Dounias, 2011; also Thomaidis & Dounias, 2012).
- **Hierarchical Forecasting:** Leverage the hierarchical structure of the dataset to reconcile forecasts across levels (e.g., bottom-up or top-down forecasts).
- **Granular Analysis:** Investigate localized trends and consumer preferences by analysing demand at the product and/or store level.

These directions can further refine forecasting practices, supporting more effective decision-making in supply chain operations.

References

- Allaire, J., Teague, C., Scheidegger, C., Xie, Y., Dervieux, C., & Woodhull, G. (2024). *Quarto* (Version 1.6). <https://doi.org/10.5281/zenodo.5960048>
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time Series Analysis* (5th ed.). Wiley.
- Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47, 1287. <https://doi.org/10.2307/1911963>
- Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on LOESS. *Journal of Official Statistics*, 6, 3–73.
- Dickey, D. A., & Fuller, W. A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica*, 49, 1057. <https://doi.org/10.2307/1912517>
- Enders, Walter. (2015). *Applied econometric time series* (4th ed., p. 485). Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- Hothorn, T., Zeileis, A., Farebrother, R. W., & Cummins, C. (2022). *Lmtest: Testing linear regression models*. <https://CRAN.R-project.org/package=lmtest>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>
- Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with Exponential Smoothing - The State Space Approach*. Springer-Verlag.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54, 159–178.
- Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303. <https://doi.org/10.1093/biomet/65.2.297>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, 38, 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>

- Maximee, C. C., Paul-Emilee, G., Arthur, P., & Renyu, Z. (2022). *Demand Prediction in Retail - A Practical Guide to Leverage Data and Predictive Analytics* (Vol. 14). Springer.
- Neusser, K. (2016). *Time Series Econometrics*. Springer.
- O'Hara-Wild, M., Hyndman, R., & Wang, E. (2024a). *Fable: Forecasting models for tidy time series*. <https://fable.tidyverts.org>
- O'Hara-Wild, M., Hyndman, R., & Wang, E. (2024b). *Feasts: Feature extraction and statistics for time series*. <http://feasts.tidyverts.org/>
- Ord, J. K., Fildes, R., & Kourentzes, N. (2017). *Principles of Business Forecasting* (2nd ed., p. 588). Wessex Press.
- Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., ... Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, 38(3), 705–871. <https://doi.org/10.1016/j.ijforecast.2021.11.001>
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71, 599–607. <https://doi.org/10.1093/biomet/71.3.599>
- Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, 34, 822–829. <https://doi.org/10.1016/j.ijforecast.2018.03.001>
- Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52, 591. <https://doi.org/10.2307/2333709>
- Sheather, S. (2009). *A Modern Approach to Regression with R*. Springer New York. <https://doi.org/10.1007/978-0-387-09608-7>
- Shmueli, Galit., & Lichtendahl, K. C. (2018). *Practical time series forecasting with R: A hands-on guide* (p. 232). Axelrod Schnall Publishers.
- Shumway, R., & Stoo, D. S. (2016). *Time Series Analysis and Its Applications with R Examples* (4th ed.). Springer.
- Sjoberg, D. D., Larmarange, J., Curry, M., Lavery, J., Whiting, K., & Zabor, E. C. (2024). *Gtsummary: Presentation-ready data summary and analytic result tables*. <https://github.com/ddsjoberg/gtsummary>
- Thomaidis, N. S., & Dounias, G. D. (2011). On detecting the optimal structure of a neural network under strong statistical features in errors. *Journal of Time Series Analysis*, 32, 204–222. <https://doi.org/10.1111/j.1467-9892.2010.00693.x>
- Thomaidis, N. S., & Dounias, G. D. (2012). A comparison of statistical tests for the adequacy of a neural network regression model. *Quantitative Finance*, 12, 437–449. <https://doi.org/>

10.1080/14697680903426573

Trapletti, A., & Hornik, K. (2024). *Tseries: Time series analysis and computational finance*.

<https://CRAN.R-project.org/package=tseries>

Vandeput, N. (2021). *Data Science for Supply Chain Forecasting* (2nd ed.). De Gruyter.

Vandeput, N. (2023). *Demand Forecasting Best Practices*. Manning.

Walmart Inc. (2024). *About Walmart - Corporate Information*. Retrieved November 25, 2024, from <https://corporate.walmart.com/about>.

Wang, E., Cook, D., Hyndman, R., & O'Hara-Wild, M. (2024). *Tsibble: Tidy temporal data frames and tools*. <https://tsibble.tidyverts.org>

Wickham, H. (2023). *Tidyverse: Easily install and load the tidyverse*. <https://tidyverse.tidyverse.org>

Zhu, H. (2024). *kableExtra: Construct complex table with kable and pipe syntax*. <http://haozhu233.github.io/kableExtra/>

A. Appendix A

A.1. Supplementary Analysis Tables

Weekly dataset with engineered features (first & last 20 rows):

week	Engineered Features				FOODS		HOBBIES		HOUSEHOLD	
	month_lbl	is_last_2wk_mth	has_snap_over2	has_event	sales	mean_price	sales	mean_price	sales	mean_price
2011 W10	Mar	0	1	1	128,823	2.85	20,447	4.45	32,444	4.97
2011 W11	Mar	0	0	1	122,932	2.83	19,755	4.57	34,910	4.97
2011 W12	Mar	1	0	0	111,979	2.80	19,833	4.56	32,701	4.93
2011 W13	Mar	1	1	0	121,015	2.81	21,793	4.53	33,737	4.94
2011 W14	Apr	0	1	0	129,467	2.87	20,335	4.65	32,223	4.88
2011 W15	Apr	0	1	0	123,350	2.88	19,998	4.75	32,777	4.89
2011 W16	Apr	1	0	1	110,804	2.95	17,479	4.87	31,408	4.91
2011 W17	Apr	1	0	1	100,337	2.93	20,724	4.65	33,011	4.92
2011 W18	May	0	1	1	114,551	2.92	19,920	4.65	30,977	4.86
2011 W19	May	0	1	0	122,750	2.94	21,045	4.63	33,742	4.95
2011 W20	May	1	0	0	103,543	2.91	19,695	4.50	31,049	4.91
2011 W21	May	1	0	0	102,581	2.90	19,025	4.68	29,546	4.83
2011 W22	Jun	0	1	1	122,670	2.94	18,392	4.83	35,572	4.87
2011 W23	Jun	0	1	1	134,305	3.00	20,081	4.58	34,076	4.88
2011 W24	Jun	0	1	1	124,439	3.00	20,189	4.45	32,886	4.90
2011 W25	Jun	1	0	0	115,785	2.96	19,761	4.48	33,494	4.87
2011 W26	Jun	1	1	0	123,605	2.93	20,616	4.60	32,809	4.82
2011 W27	Jul	0	1	1	138,547	3.00	20,127	4.49	35,988	4.91
2011 W28	Jul	0	1	0	132,820	3.01	19,633	4.44	34,215	4.89
2011 W29	Jul	1	0	0	122,688	2.99	20,348	4.48	34,238	4.82
2016 W05	Feb	0	1	1	196,176	3.17	27,960	5.63	69,396	4.76
2016 W06	Feb	0	1	1	207,012	3.19	27,188	5.63	71,328	4.80
2016 W07	Feb	1	0	1	198,657	3.17	27,506	5.67	74,002	4.80
2016 W08	Feb	1	0	0	178,807	3.15	26,836	5.61	70,800	4.78
2016 W09	Mar	0	1	0	199,561	3.16	27,853	5.60	75,163	4.80
2016 W10	Mar	0	1	0	202,687	3.17	25,958	5.68	69,962	4.75
2016 W11	Mar	0	0	1	193,108	3.16	26,519	5.56	67,798	4.77
2016 W12	Mar	1	0	1	195,565	3.13	25,605	5.59	66,288	4.78
2016 W13	Mar	1	1	0	189,164	3.15	28,311	5.57	74,096	4.80
2016 W14	Apr	0	1	0	211,390	3.19	27,290	5.48	72,087	4.83
2016 W15	Apr	0	1	0	201,045	3.19	27,683	5.35	67,281	4.83
2016 W16	Apr	1	0	0	189,333	3.16	27,490	5.54	68,456	4.84
2016 W17	Apr	1	0	1	191,019	3.14	29,056	5.47	69,739	4.83
2016 W18	May	0	1	1	213,122	3.17	28,886	5.44	69,504	4.80
2016 W19	May	0	1	0	225,812	3.16	28,786	5.54	72,234	4.85
2016 W20	May	1	0	0	202,535	3.12	29,128	5.52	71,943	4.85
2016 W21	May	1	0	0	190,609	3.09	28,083	5.52	67,764	4.84
2016 W22	Jun	0	1	1	207,344	3.09	29,921	5.51	78,291	4.88
2016 W23	Jun	0	1	1	219,853	3.13	28,575	5.58	71,678	4.85
2016 W24	Jun	0	1	1	209,304	3.12	28,739	5.51	68,981	4.87

Weekly dataset with lags of sales (first & last 30 rows):

week	FOODS					HOBBIES					HOUSEHOLD				
	sales	lag_1	lag_2	lag_3	lag_4	sales	lag_1	lag_2	lag_3	lag_4	sales	lag_1	lag_2	lag_3	lag_4
2011 W10	128,823	130,587	114,126	120,641	141,740	20,447	14,678	19,392	18,410	20,257	32,444	34,018	33,127	33,277	31,718
2011 W11	122,932	128,823	130,587	114,126	120,641	19,755	20,447	14,678	19,392	18,410	34,910	32,444	34,018	33,127	33,277
2011 W12	111,979	122,932	128,823	130,587	114,126	19,833	19,755	20,447	14,678	19,392	32,701	34,910	32,444	34,018	33,127
2011 W13	121,015	111,979	122,932	128,823	130,587	21,793	19,833	19,755	20,447	14,678	33,737	32,701	34,910	32,444	34,018
2011 W14	129,467	121,015	111,979	122,932	128,823	20,335	21,793	19,833	19,755	20,447	32,223	33,737	32,701	34,910	32,444
2011 W15	123,350	129,467	121,015	111,979	122,932	19,998	20,335	21,793	19,833	19,755	32,777	32,223	33,737	32,701	34,910
2011 W16	110,804	123,350	129,467	121,015	111,979	17,479	19,998	20,335	21,793	19,833	31,408	32,777	32,223	33,737	32,701
2011 W17	100,337	110,804	123,350	129,467	121,015	20,724	17,479	19,998	20,335	21,793	33,011	31,408	32,777	32,223	33,737
2011 W18	114,551	100,337	110,804	123,350	129,467	19,920	20,724	17,479	19,998	20,335	30,977	33,011	31,408	32,777	32,223
2011 W19	122,750	114,551	100,337	110,804	123,350	21,045	19,920	20,724	17,479	19,998	33,742	30,977	33,011	31,408	32,777
2011 W20	103,543	122,750	114,551	100,337	110,804	19,695	21,045	19,920	20,724	17,479	31,049	33,742	30,977	33,011	31,408
2011 W21	102,581	103,543	122,750	114,551	100,337	19,025	19,695	21,045	19,920	20,724	29,546	31,049	33,742	30,977	33,011
2011 W22	122,670	102,581	103,543	122,750	114,551	18,392	19,025	19,695	21,045	19,920	35,572	29,546	31,049	33,742	30,977
2011 W23	134,305	122,670	102,581	103,543	122,750	20,081	18,392	19,025	19,695	21,045	34,076	35,572	29,546	31,049	33,742
2011 W24	124,439	134,305	122,670	102,581	103,543	20,189	20,081	18,392	19,025	19,695	32,886	34,076	35,572	29,546	31,049
2011 W25	115,785	124,439	134,305	122,670	102,581	19,761	20,189	20,081	18,392	19,025	33,494	32,886	34,076	35,572	29,546
2011 W26	123,605	115,785	124,439	134,305	122,670	20,616	19,761	20,189	20,081	18,392	32,809	33,494	32,886	34,076	35,572
2011 W27	138,547	123,605	115,785	124,439	134,305	20,127	20,616	19,761	20,189	20,081	35,988	32,809	33,494	32,886	34,076
2011 W28	132,820	138,547	123,605	115,785	124,439	19,633	20,127	20,616	19,761	20,189	34,215	35,988	32,809	33,494	32,886
2011 W29	122,688	132,820	138,547	123,605	115,785	20,348	19,633	20,127	20,616	19,761	34,238	34,215	35,988	32,809	33,494
2011 W30	114,079	122,688	132,820	138,547	123,605	19,491	20,348	19,633	20,127	20,616	34,439	34,238	34,215	35,988	32,809
2011 W31	135,733	114,079	122,688	132,820	138,547	20,314	19,491	20,348	19,633	20,127	37,544	34,439	34,238	34,215	35,988
2011 W32	141,460	135,733	114,079	122,688	132,820	19,610	20,314	19,491	20,348	19,633	36,239	37,544	34,439	34,238	34,215
2011 W33	136,544	141,460	135,733	114,079	122,688	19,517	19,610	20,314	19,491	20,348	38,008	36,239	37,544	34,439	34,238
2011 W34	122,144	136,544	141,460	135,733	114,079	19,019	19,517	19,610	20,314	19,491	35,404	38,008	36,239	37,544	34,439
2011 W35	126,182	122,144	136,544	141,460	135,733	18,031	19,019	19,517	19,610	20,314	36,409	35,404	38,008	36,239	37,544
2011 W36	142,491	126,182	122,144	136,544	141,460	19,018	18,031	19,019	19,517	19,610	39,184	36,409	35,404	38,008	36,239
2011 W37	136,701	142,491	126,182	122,144	136,544	17,851	19,018	18,031	19,019	19,517	37,241	39,184	36,409	35,404	38,008
2011 W38	128,815	136,701	142,491	126,182	122,144	17,949	17,851	19,018	18,031	19,019	36,333	37,241	39,184	36,409	35,404
2011 W39	139,181	128,815	136,701	142,491	126,182	19,976	17,949	17,851	19,018	18,031	39,005	36,333	37,241	39,184	36,409
2015 W48	150,214	163,600	178,766	177,296	165,798	24,903	28,916	27,873	29,411	29,827	57,575	62,993	63,522	69,897	66,536
2015 W49	172,158	150,214	163,600	178,766	177,296	30,690	24,903	28,916	27,873	29,411	69,374	57,575	62,993	63,522	69,897
2015 W50	171,224	172,158	150,214	163,600	178,766	28,512	30,690	24,903	28,916	27,873	61,885	69,374	57,575	62,993	63,522
2015 W51	172,620	171,224	172,158	150,214	163,600	29,474	28,512	30,690	24,903	28,916	62,390	61,885	69,374	57,575	62,993
2015 W52	140,943	172,620	171,224	172,158	150,214	24,729	29,474	28,512	30,690	24,903	55,210	62,390	61,885	69,374	57,575
2015 W53	166,308	140,943	172,620	171,224	172,158	26,913	24,729	29,474	28,512	30,690	67,550	55,210	62,390	61,885	69,374
2016 W01	188,614	166,308	140,943	172,620	171,224	26,622	26,913	24,729	29,474	28,512	64,577	67,550	55,210	62,390	61,885
2016 W02	192,196	188,614	166,308	140,943	172,620	25,274	26,622	26,913	24,729	29,474	62,194	64,577	67,550	55,210	62,390
2016 W03	177,317	192,196	188,614	166,308	140,943	25,832	25,274	26,622	26,913	24,729	64,708	62,194	64,577	67,550	55,210
2016 W04	171,099	177,317	192,196	188,614	166,308	27,832	25,832	25,274	26,622	26,913	67,293	64,708	62,194	64,577	67,550
2016 W05	196,176	171,099	177,317	192,196	188,614	27,960	27,832	25,274	26,622	26,913	69,396	67,293	64,708	62,194	64,577
2016 W06	207,012	196,176	171,099	177,317	192,196	27,188	27,960	27,832	25,274	26,913	71,328	69,396	67,293	64,708	62,194
2016 W07	198,657	207,012	196,176	171,099	177,317	27,506	27,188	27,960	27,832	25,832	74,002	71,328	69,396	67,293	64,708
2016 W08	178,807	198,657	207,012	196,176	171,099	26,836	27,506	27,188	27,960	27,832	70,800	74,002	71,328	69,396	67,293
2016 W09	199,561	178,807	198,657	207,012	196,176	27,853	26,836	27,506	27,188	27,960	75,163	70,800	74,002	71,328	69,396
2016 W10	202,687	199,561	178,807	198,657	207,012	25,958	27,853	26,836	27,506	27,188	69,962	75,163	70,800	74,002	71,328
2016 W11	193,108	202,687	199,561	178,807	198,657	26,519	25,958	27,853	26,836	27,506	67,798	69,962	75,163	70,800	74,002
2016 W12	195,565	193,108	202,687	199,561	178,807	25,605	26,519	25,958	27,853	26,836	66,288	67,798	69,962	75,163	70,800
2016 W13	189,164	195,565	193,108	202,687	199,561	28,311	25,605	26,519	25,958	27,853	74,096	66,288	67,798	69,962	75,163
2016 W14	211,390	189,164	195,565	193,108	202,687	27,290	28,311	25,605	26,519	25,958	72,087	74,096	66,288	67,798	69,962
2016 W15	201,045	211,390	189,164	195,565	193,108	27,683	27,290	28,311	25,605	26,519	67,281	72,087	74,096	66,288	67,798
2016 W16	189,333	201,045	211,390	189,164	195,565	27,490	27,683	27,290	28,311	25,605	68,456	67,281	72,087	74,096	66,288
2016 W17	191,019	189,333	201,045	211,390	189,164	29,056	27,490	27,683	27,290	28,311	69,739	68,456	67,281	72,087	74,096
2016 W18	213,122	191,019	189,333	201,045	211,390	28,886	29,056	27,490	27,683	27,290	69,504	69,739	68,456	67,281	72,087
2016 W19	225,812	213,122	191,019	189,333	201,045	28,786	28,886	29,056	27,490	27,683	72,234	69,504	69,739	68,456	67,281
2016 W20	202,535	225,812	213,122	191,019	189,333	29,128	28,786	28,886	29,056	27,490	71,943	72,234	69,504	69,739	68,456
2016 W21	190,609	202,535	225,812	213,122	191,019	28,083	29,128	28,786	28,886	29,056	67,764	71,943	72,234	69,504	69,739
2016 W22	207,344	190,609	202,535	225,812	213,122	29,921	28,083	29,128	28,786	28,886	78,291	67,764	71,943	72,234	69,504
2016 W23	219,853	207,344	190,609	202,535	225,812	28,575	29,921	28,083	29,128	28,786	71,678	78,291	67,764	71,943	72,234
2016 W24	209,304	219,853	207,344	190,609	202,535	28,739	28,575	29,921	28,083	29,128	68,981	71,678	78,291	67,764	71,943

STL decomposition results (first 40 rows):

FOODS					
week	sales	trend	season_year	remainder	season_adjust
2011 W10	128,823	108,569.7	8,697.77	11,555.50	120,125.2
2011 W11	122,932	109,502.4	8,621.36	4,808.28	114,310.6
2011 W12	111,979	110,435.0	-8,471.31	10,015.32	120,450.3
2011 W13	121,015	111,367.6	-8,996.25	18,643.63	130,011.2
2011 W14	129,467	112,300.2	9,048.82	8,117.92	120,418.2
2011 W15	123,350	113,232.9	8,901.80	1,215.31	114,448.2
2011 W16	110,804	114,165.5	-5,223.25	1,861.73	116,027.2
2011 W17	100,337	115,098.1	-14,510.74	-250.41	114,847.7
2011 W18	114,551	116,030.8	-1,765.15	285.36	116,316.1
2011 W19	122,750	116,974.1	6,659.12	-883.22	116,090.9
2011 W20	103,543	117,917.4	2,586.47	-16,960.88	100,956.5
2011 W21	102,581	118,860.7	-13,346.13	-2,933.59	115,927.1
2011 W22	122,670	119,804.0	-11,166.78	14,032.74	133,836.8
2011 W23	134,305	120,747.4	15,279.52	-1,721.87	119,025.5
2011 W24	124,439	121,690.7	17,121.49	-14,373.15	107,317.5
2011 W25	115,785	122,634.0	2,539.71	-9,388.68	113,245.3
2011 W26	123,605	123,577.3	-3,993.55	4,021.26	127,598.6
2011 W27	138,547	124,516.8	22,356.81	-8,326.59	116,190.2
2011 W28	132,820	125,456.3	20,908.93	-13,545.21	111,911.1
2011 W29	122,688	126,395.8	4,462.69	-8,170.47	118,225.3
2011 W30	114,079	127,335.3	-8,179.34	-5,076.93	122,258.3
2011 W31	135,733	128,274.8	7,261.55	196.68	128,471.4
2011 W32	141,460	129,214.3	20,781.48	-8,535.74	120,678.5
2011 W33	136,544	130,153.8	15,676.55	-9,286.31	120,867.4
2011 W34	122,144	131,093.3	-1,280.55	-7,668.71	123,424.6
2011 W35	126,182	132,026.5	-4,966.43	-878.08	131,148.4
2011 W36	142,491	132,959.8	13,855.27	-4,324.03	128,635.7
2011 W37	136,701	133,893.0	15,457.31	-12,649.32	121,243.7
2011 W38	128,815	134,826.3	-1,838.32	-4,172.94	130,653.3
2011 W39	139,181	135,759.5	-6,861.77	10,283.26	146,042.8
2011 W40	154,718	136,692.8	11,077.97	6,947.27	143,640.0
2011 W41	156,503	137,626.0	14,501.30	4,375.69	142,001.7
2011 W42	133,759	138,559.3	-2,222.62	-2,577.64	135,981.6
2011 W43	127,534	139,540.7	-16,767.55	4,760.81	144,301.5
2011 W44	145,009	140,522.2	-6,444.65	10,931.42	151,453.6
2011 W45	150,421	141,503.7	5,788.47	3,128.82	144,632.5
2011 W46	138,248	142,485.2	-2,560.12	-1,677.07	140,808.1
2011 W47	127,715	143,466.7	-20,091.39	4,339.71	147,806.4
2011 W48	137,657	144,448.2	-26,966.41	20,175.25	164,623.4
2011 W49	154,699	145,429.6	4,960.63	4,308.72	149,738.4

HOBBIES					
week	sales	trend	season_year	remainder	season_adjust
2011 W10	20,447	19,705.80	348.11	393.09	20,098.89
2011 W11	19,755	19,701.54	-340.77	394.23	20,095.77
2011 W12	19,833	19,697.27	-481.05	616.78	20,314.05
2011 W13	21,793	19,693.01	-429.86	2,529.85	22,222.86
2011 W14	20,335	19,688.75	645.10	1.16	19,689.90
2011 W15	19,998	19,684.48	473.38	-159.87	19,524.62
2011 W16	17,479	19,680.22	-765.38	-1,435.84	18,244.38
2011 W17	20,724	19,675.95	88.01	960.04	20,635.99
2011 W18	19,920	19,671.69	476.40	-228.09	19,443.60
2011 W19	21,045	19,668.57	561.69	814.74	20,483.31
2011 W20	19,695	19,665.46	412.86	-383.31	19,282.14
2011 W21	19,025	19,662.34	261.83	-899.17	18,763.17
2011 W22	18,392	19,659.22	619.67	-1,886.89	17,772.33
2011 W23	20,081	19,656.10	1,369.40	-944.50	18,711.60
2011 W24	20,189	19,652.98	235.29	300.72	19,953.71
2011 W25	19,761	19,649.86	478.54	-367.41	19,282.46
2011 W26	20,616	19,646.75	158.07	811.18	20,457.93
2011 W27	20,127	19,644.89	842.23	-360.12	19,284.77
2011 W28	19,633	19,643.04	-126.12	116.08	19,759.12
2011 W29	20,348	19,641.19	457.20	249.61	19,890.80
2011 W30	19,491	19,639.33	-1,291.61	1,143.28	20,782.61
2011 W31	20,314	19,637.48	608.52	68.00	19,705.48
2011 W32	19,610	19,635.63	-455.03	429.40	20,065.03
2011 W33	19,517	19,633.77	-784.26	667.49	20,301.26
2011 W34	19,019	19,631.92	-801.28	188.36	19,820.28
2011 W35	18,031	19,631.75	-1,753.00	152.25	19,784.00
2011 W36	19,018	19,631.57	-357.50	-256.07	19,375.50
2011 W37	17,851	19,631.39	-1,229.12	-551.27	19,080.12
2011 W38	17,949	19,631.22	-1,402.59	-279.63	19,351.59
2011 W39	19,976	19,631.04	-377.62	722.58	20,353.62
2011 W40	19,362	19,630.87	814.23	-1,083.09	18,547.77
2011 W41	19,751	19,630.69	185.22	-64.91	19,565.78
2011 W42	19,258	19,630.51	681.87	-1,054.38	18,576.13
2011 W43	20,208	19,633.59	1,066.85	-492.44	19,141.15
2011 W44	21,092	19,636.67	1,811.24	-355.90	19,280.76
2011 W45	18,836	19,639.75	123.43	-927.18	18,712.57
2011 W46	18,998	19,642.83	202.49	-847.31	18,795.51
2011 W47	16,595	19,645.90	1,025.09	-4,075.99	15,569.91
2011 W48	21,482	19,648.98	-2,887.27	4,720.28	24,369.27
2011 W49	19,814	19,652.06	2,140.73	-1,978.79	17,673.27

HOUSEHOLD					
week	sales	trend	season_year	remainder	season_adjust
2011 W10	32,444	30,124.29	2,197.78	121.93	30,246.22
2011 W11	34,910	30,316.78	3,057.24	1,535.98	31,852.76
2011 W12	32,701	30,509.26	193.83	1,997.91	32,507.17
2011 W13	33,737	30,701.75	-1,226.15	4,261.40	34,963.15
2011 W14	32,223	30,894.24	2,902.50	-1,573.74	29,320.50
2011 W15	32,777	31,086.73	1,995.94	-305.67	30,781.06
2011 W16	31,408	31,279.22	-641.61	770.40	32,049.61
2011 W17	33,011	31,471.71	-835.56	2,374.85	33,846.56
2011 W18	30,977	31,664.19	-60.97	-626.22	31,037.97
2011 W19	33,742	31,859.08	-1,413.95	3,296.87	35,155.95
2011 W20	31,049	32,053.96	1,019.46	-2,024.43	30,029.54
2011 W21	29,546	32,248.85	-1,870.96	-831.89	31,416.96
2011 W22	35,572	32,443.74	2,017.32	1,110.95	33,554.68
2011 W23	34,076	32,638.62	3,483.26	-2,045.89	30,592.74
2011 W24	32,886	32,833.51	-2.74	55.24	32,888.74
2011 W25	33,494	33,028.39	-80.12	545.73	33,574.12
2011 W26	32,809	33,223.28	-386.70	-27.58	33,195.70
2011 W27	35,988	33,418.44	3,835.32	-1,265.76	32,152.68
2011 W28	34,215	33,613.60	2,023.80	-1,422.40	32,191.20
2011 W29	34,238	33,808.76	1,755.23	-1,325.99	32,482.77
2011 W30	34,439	34,003.92	271.46	163.62	34,167.54
2011 W31	37,544	34,199.08	4,518.36	-1,173.45	33,025.64
2011 W32	36,239	34,394.24	3,607.36	-1,762.61	32,631.64
2011 W33	38,008	34,589.40	4,108.84	-690.24	33,899.16
2011 W34	35,404	34,784.57	4,488.01	-3,868.58	30,915.99
2011 W35	36,409	34,979.85	3,467.58	-2,038.43	32,941.42
2011 W36	39,184	35,175.13	4,701.79	-692.92	34,482.21
2011 W37	37,241	35,370.42	2,732.39	-861.80	34,508.61
2011 W38	36,333	35,565.70	1,527.10	-759.80	34,805.90
2011 W39	39,005	35,760.98	1,510.79	1,733.23	37,494.21
2011 W40	36,323	35,956.26	6,234.86	-5,868.13	30,088.14
2011 W41	38,542	36,151.55	2,088.34	302.11	36,453.66
2011 W42	36,777	36,346.83	-172.65	602.82	36,949.65
2011 W43	36,287	36,556.26	-1,946.71	1,677.45	38,233.71
2011 W44	38,713	36,765.69	712.33	1,234.98	38,000.67
2011 W45	36,473	36,975.12	-41.35	-460.77	36,514.35
2011 W46	36,075	37,184.56	-3,682.60	2,573.05	39,757.60
2011 W47	32,556	37,393.99	-5,116.64	278.66	37,672.64
2011 W48	35,114	37,603.42	-7,949.47	5,460.05	43,063.47
2011 W49	33,741	37,812.85	-2,240.00	-1,831.85	35,981.00

A.2. Supplementary Modeling Tables

Regression results summary tables:

FOODS								
Variable	Initial Model				Refined Model			
	Estimate	95% CI	P-value	GVIF	Estimate	95% CI	P-value	GVIF
(Intercept)	319,285	249,340, 389,231	<0.001		32,234	-12,476, 76,943	0.157	
Trend	243	215, 272	<0.001	1.9	24	0.47, 47	0.046	4.7
Month dummies				1.3				1.5
Jan	—	—			—	—		
Feb	4,802	-3,702, 13,306	0.267		-2,553	-7,124, 2,017	0.272	
Mar	-312	-8,171, 7,547	0.938		-5,908	-10,142, -1,674	0.006	
Apr	260	-7,920, 8,441	0.950		-5,131	-9,515, -747	0.022	
May	-1,067	-9,273, 7,139	0.798		-4,870	-9,266, -474	0.030	
Jun	4,645	-3,270, 12,559	0.249		-406	-4,626, 3,814	0.850	
Jul	8,380	-187, 16,947	0.055		-3,410	-8,073, 1,252	0.151	
Aug	7,356	-1,327, 16,039	0.096		-3,058	-7,775, 1,660	0.203	
Sept	5,003	-3,295, 13,301	0.236		-5,014	-9,523, -506	0.029	
Oct	681	-7,894, 9,257	0.876		-8,380	-13,042, -3,718	<0.001	
Nov	-5,630	-14,021, 2,762	0.188		-9,530	-14,069, -4,991	<0.001	
Dec	-8,292	-16,423, -160	0.046		-7,227	-11,566, -2,887	0.001	
End-Week dummy				2.0				2.7
0	—	—			—	—		
1	-3,023	-7,616, 1,570	0.196		-2,578	-5,402, 245	0.073	
SNAP dummy				2.0				2.3
0	—	—			—	—		
1	18,718	14,101, 23,335	<0.001		9,932	7,332, 12,532	<0.001	
Event dummy				1.2				1.2
0	—	—			—	—		
1	-269	-3,775, 3,237	0.880		1,123	-742, 2,988	0.237	
Price	-65,761	-90,005, -41,516	<0.001	2.0	-6,442	-20,462, 7,579	0.366	2.4
Lag 1					0.44	0.38, 0.50	<0.001	2.8
Lag 4					0.46	0.38, 0.54	<0.001	4.6
R ²	0.674				0.909			
Adjusted R ²	0.654				0.902			
Statistic	33.4				142			
df	16				18			
p-value	<0.001				<0.001			
AIC	6,057				5,709			

Abbreviations: CI = Confidence Interval, GVIF = Generalized Variance Inflation Factor

HOBIES								
Variable	Initial Model				Refined Model			
	Estimate	95% CI	P-value	GVIF	Estimate	95% CI	P-value	GVIF
(Intercept)	34,307	29,798, 38,816	<0.001		14,966	10,160, 19,773	<0.001	
Trend	54	49, 59	<0.001	3.3	20	14, 27	<0.001	9.7
Month dummies				1.3				1.7
Jan	—	—			—	—		
Feb	1,491	352, 2,630	0.010		467	-475, 1,409	0.330	
Mar	1,344	293, 2,395	0.012		-52	-923, 818	0.906	
Apr	1,706	611, 2,801	0.002		231	-677, 1,140	0.616	
May	943	-158, 2,045	0.093		-371	-1,278, 535	0.421	
Jun	1,954	893, 3,016	<0.001		328	-564, 1,219	0.470	
Jul	1,486	342, 2,630	0.011		-237	-1,193, 720	0.626	
Aug	565	-592, 1,721	0.337		-583	-1,528, 361	0.225	
Sept	-40	-1,143, 1,063	0.943		-511	-1,397, 376	0.258	
Oct	497	-664, 1,659	0.400		-70	-1,006, 867	0.884	
Nov	359	-776, 1,495	0.533		-814	-1,742, 114	0.085	
Dec	-0.84	-1,091, 1,089	0.999		-718	-1,599, 162	0.109	
End-Week dummy				2.0				2.0
0	—	—			—	—		
1	-322	-936, 292	0.303		-412	-907, 84	0.103	
SNAP dummy				2.0				2.0
0	—	—			—	—		
1	410	-199, 1,020	0.186		410	-80, 900	0.101	
Event dummy				1.2				1.2
0	—	—			—	—		
1	-292	-762, 178	0.222		-316	-700, 68	0.106	
Price	-3,850	-4,792, -2,907	<0.001	3.4	-1,967	-2,782, -1,152	<0.001	3.9
Lag 1					0.22	0.10, 0.34	<0.001	6.2
Lag 2					0.31	0.20, 0.43	<0.001	5.7
Lag 3					0.14	0.03, 0.26	0.014	5.7
R ²	0.765				0.851			
Adjusted R ²	0.750				0.840			
Statistic	52.6				77.1			
df	16				19			
p-value	<0.001				<0.001			
AIC	4,946				4,826			

Abbreviations: CI = Confidence Interval, GVIF = Generalized Variance Inflation Factor

HOUSEHOLD								
Variable	Initial Model				Refined Model			
	Estimate	95% CI	P-value	GVI	Estimate	95% CI	P-value	GVI
(Intercept)	175,398	144,971, 205,825	<0.001		24,892	-1,891, 51,676	0.068	
Trend	122	115, 128	<0.001	1.2	17	4.7, 30	0.007	9.6
Month dummies				1.4				3.3
Jan	—	—			—	—		
Feb	8,773	6,223, 11,322	<0.001		3,901	2,055, 5,746	<0.001	
Mar	8,916	6,515, 11,317	<0.001		-898	-2,875, 1,080	0.372	
Apr	7,228	4,753, 9,702	<0.001		-674	-2,585, 1,236	0.487	
May	4,849	2,385, 7,312	<0.001		-1,119	-2,935, 696	0.226	
Jun	7,048	4,697, 9,399	<0.001		793	-954, 2,540	0.372	
Jul	6,897	4,371, 9,424	<0.001		79	-1,802, 1,961	0.934	
Aug	8,746	6,193, 11,300	<0.001		1,382	-542, 3,305	0.159	
Sept	9,034	6,594, 11,473	<0.001		-334	-2,304, 1,635	0.738	
Oct	6,540	4,011, 9,069	<0.001		-1,722	-3,692, 249	0.087	
Nov	3,101	618, 5,584	0.015		-2,445	-4,264, -626	0.009	
Dec	-120	-2,528, 2,288	0.922		-2,311	-3,999, -624	0.007	
End-Week dummy				2.0				2.1
0	—	—			—	—		
1	-1,051	-2,407, 305	0.128		-1,285	-2,219, -352	0.007	
SNAP dummy				2.0				2.1
0	—	—			—	—		
1	1,205	-140, 2,551	0.079		1,702	759, 2,645	<0.001	
Event dummy				1.2				1.2
0	—	—			—	—		
1	-173	-1,209, 862	0.742		-59	-771, 653	0.871	
Price	-30,352	-36,702, -24,002	<0.001	1.3	-4,016	-9,257, 1,226	0.133	2.0
Lag 1					0.16	0.04, 0.27	0.009	18
Lag 2					0.39	0.28, 0.50	<0.001	17
Lag 3					0.31	0.19, 0.42	<0.001	18
R ²	0.891				0.950			
Adjusted R ²	0.884				0.947			
Statistic	132				257			
df	16				19			
p-value	<0.001				<0.001			
AIC	5,383				5,173			

Abbreviations: CI = Confidence Interval, GVI = Generalized Variance Inflation Factor

Out-of-sample results table:

Week	FOODS				HOBBIES				HOUSEHOLD			
	Sales	SES	HW	LM	Sales	SES	HW	LM	Sales	SES	HW	LM
2015 W23	187,633	162,420.2	191,544.3	176,460.4	27,683	26,663.13	27,975.58	27,174.53	69,496	65,404.02	67,007.07	67,785.70
2015 W24	192,488	162,420.2	192,485.8	188,831.7	25,266	26,663.13	27,143.38	27,439.27	65,229	65,404.02	62,081.48	69,200.85
2015 W25	179,213	162,420.2	175,031.9	168,730.1	26,060	26,663.13	27,814.51	27,088.76	63,405	65,404.02	66,029.42	67,808.51
2015 W26	162,650	162,420.2	163,856.6	164,309.9	25,483	26,663.13	27,386.06	26,831.45	64,886	65,404.02	63,946.57	66,160.35
2015 W27	183,195	162,420.2	198,883.7	182,736.2	27,730	26,663.13	27,784.06	25,495.97	70,278	65,404.02	69,785.66	65,827.41
2015 W28	196,107	162,420.2	194,767.0	193,909.3	27,272	26,663.13	26,485.90	26,450.38	66,667	65,404.02	67,526.21	67,218.30
2015 W29	189,667	162,420.2	171,740.0	191,101.0	27,161	26,663.13	27,595.15	26,785.83	69,925	65,404.02	68,197.50	67,675.52
2015 W30	170,179	162,420.2	157,686.9	169,923.7	26,090	26,663.13	25,956.83	26,805.31	68,244	65,404.02	65,721.65	66,970.94
2015 W31	177,095	162,420.2	173,041.6	173,977.5	27,389	26,663.13	27,902.50	26,119.21	71,946	65,404.02	71,241.60	70,539.68
2015 W32	198,895	162,420.2	187,253.7	193,506.1	27,645	26,663.13	25,878.44	26,299.93	72,654	65,404.02	69,028.66	72,431.46
2015 W33	190,083	162,420.2	187,483.7	197,941.3	26,877	26,663.13	25,845.00	26,804.56	70,957	65,404.02	70,284.68	72,248.17
2015 W34	170,866	162,420.2	172,831.4	174,270.0	26,847	26,663.13	25,287.56	26,636.19	72,506	65,404.02	72,007.60	72,067.69
2015 W35	163,579	162,420.2	164,966.8	168,738.4	26,881	26,663.13	25,145.47	26,587.95	69,975	65,404.02	70,684.03	71,709.95
2015 W36	182,777	162,420.2	191,210.1	186,783.9	27,533	26,663.13	24,783.94	26,391.39	71,744	65,404.02	71,922.13	72,547.63
2015 W37	199,434	162,420.2	184,122.9	190,424.5	26,873	26,663.13	25,320.82	26,683.46	71,148	65,404.02	69,322.70	72,147.11
2015 W38	184,838	162,420.2	160,131.0	176,155.2	27,238	26,663.13	25,451.00	26,688.19	68,010	65,404.02	69,487.75	69,516.86
2015 W39	174,899	162,420.2	152,541.7	166,270.6	26,427	26,663.13	25,545.05	26,818.34	67,730	65,404.02	67,503.19	69,020.56
2015 W40	185,259	162,420.2	175,930.6	180,011.7	28,675	26,663.13	27,940.84	27,296.90	71,848	65,404.02	76,412.27	68,818.92
2015 W41	200,451	162,420.2	181,015.1	192,886.6	27,168	26,663.13	27,290.42	27,869.38	69,502	65,404.02	71,035.67	68,802.66
2015 W42	188,906	162,420.2	167,343.4	190,370.3	28,451	26,663.13	27,671.76	28,059.69	66,364	65,404.02	69,074.56	68,507.15
2015 W43	170,327	162,420.2	152,339.0	170,406.9	29,022	26,663.13	27,547.76	28,609.61	63,444	65,404.02	67,564.24	66,713.19
2015 W44	165,798	162,420.2	166,419.7	170,453.9	29,827	26,663.13	28,329.39	27,674.61	66,536	65,404.02	68,548.72	65,222.27
2015 W45	177,296	162,420.2	184,713.4	186,300.3	29,411	26,663.13	26,651.99	28,008.80	69,897	65,404.02	71,462.70	64,837.70
2015 W46	178,766	162,420.2	177,676.9	185,289.9	27,873	26,663.13	26,780.18	28,383.82	63,522	65,404.02	64,945.92	66,270.58
2015 W47	163,600	162,420.2	157,453.7	164,918.2	28,916	26,663.13	26,617.02	27,817.87	62,993	65,404.02	66,043.92	64,531.20
2015 W48	150,214	162,420.2	152,259.6	156,197.4	24,903	26,663.13	24,370.61	27,789.75	57,575	65,404.02	62,610.74	62,314.11
2015 W49	172,158	162,420.2	182,904.3	169,164.2	30,690	26,663.13	28,880.15	26,846.36	69,374	65,404.02	69,785.42	61,913.54
2015 W50	171,224	162,420.2	181,901.4	179,205.0	28,512	26,663.13	26,775.02	27,702.40	61,885	65,404.02	62,561.02	61,800.89
2015 W51	172,620	162,420.2	173,160.3	161,069.5	29,474	26,663.13	27,455.40	27,841.72	62,390	65,404.02	62,484.77	63,262.87
2015 W52	140,943	162,420.2	146,221.1	153,331.0	24,729	26,663.13	23,476.72	28,040.59	55,210	65,404.02	58,808.19	61,190.15
2015 W53	166,308	162,420.2	183,924.3	161,127.7	26,913	26,663.13	26,767.73	26,582.53	67,550	65,404.02	67,825.25	59,131.64
2016 W01	188,614	162,420.2	197,301.8	179,016.6	26,622	26,663.13	25,816.59	27,114.11	64,577	65,404.02	62,341.42	62,988.44
2016 W02	192,196	162,420.2	185,405.6	189,172.8	25,274	26,663.13	26,731.58	26,966.29	62,194	65,404.02	68,603.11	66,294.53
2016 W03	177,317	162,420.2	164,409.0	163,022.4	25,832	26,663.13	27,712.57	27,027.89	64,708	65,404.02	67,896.03	64,757.23
2016 W04	171,099	162,420.2	170,732.3	169,652.0	27,832	26,663.13	29,137.74	26,945.14	67,293	65,404.02	71,457.77	63,469.98
2016 W05	196,176	162,420.2	196,929.7	186,944.9	27,960	26,663.13	29,831.70	27,611.66	69,396	65,404.02	76,911.16	70,766.68
2016 W06	207,012	162,420.2	193,471.3	199,387.5	27,188	26,663.13	29,439.58	28,246.31	71,328	65,404.02	74,149.71	73,036.09
2016 W07	198,657	162,420.2	167,909.2	184,864.5	27,506	26,663.13	28,778.24	28,287.54	74,002	65,404.02	77,098.88	72,357.02
2016 W08	178,807	162,420.2	169,688.0	178,165.7	26,836	26,663.13	29,635.76	28,334.08	70,800	65,404.02	79,114.99	74,345.22
2016 W09	199,561	162,420.2	195,189.0	191,474.4	27,853	26,663.13	28,711.49	27,750.94	75,163	65,404.02	74,640.24	72,828.70
2016 W10	202,687	162,420.2	198,462.5	205,782.8	25,958	26,663.13	28,458.63	27,779.58	69,962	65,404.02	79,363.97	72,946.86
2016 W11	193,108	162,420.2	179,947.0	192,835.3	26,519	26,663.13	28,115.21	27,672.70	67,798	65,404.02	74,624.78	72,008.66
2016 W12	195,565	162,420.2	176,461.1	177,159.1	25,605	26,663.13	27,769.96	27,443.17	66,288	65,404.02	72,444.92	68,593.77
2016 W13	189,164	162,420.2	200,957.5	198,795.5	28,311	26,663.13	29,368.12	27,126.79	74,096	65,404.02	77,348.39	67,071.96
2016 W14	211,390	162,420.2	202,925.6	199,611.6	27,290	26,663.13	29,380.23	28,193.18	72,087	65,404.02	78,152.98	69,690.82
2016 W15	201,045	162,420.2	192,332.0	204,143.0	27,683	26,663.13	27,778.18	28,750.14	67,281	65,404.02	74,567.16	72,506.10
2016 W16	189,333	162,420.2	175,232.2	189,487.4	27,490	26,663.13	29,177.54	28,635.60	68,456	65,404.02	74,043.35	69,782.83
2016 W17	191,019	162,420.2	187,450.5	181,274.7	29,056	26,663.13	28,454.20	28,687.26	69,739	65,404.02	74,212.14	67,329.95
2016 W18	213,122	162,420.2	193,670.7	204,270.9	28,886	26,663.13	28,794.20	28,551.98	69,504	65,404.02	72,813.44	69,217.44
2016 W19	225,812	162,420.2	193,953.0	208,360.0	28,786	26,663.13	28,474.85	28,714.13	72,234	65,404.02	77,012.08	70,059.24
2016 W20	202,535	162,420.2	177,083.9	196,235.1	29,128	26,663.13	28,309.62	28,914.65	71,943	65,404.02	72,912.77	68,150.10
2016 W21	190,609	162,420.2	184,886.0	187,247.7	28,083	26,663.13	28,771.17	28,982.07	67,764	65,404.02	76,415.79	69,300.70
2016 W22	207,344	162,420.2	223,015.1	211,597.8	29,921	26,663.13	29,825.13	29,638.71	78,291	65,404.02	76,779.97	74,069.97
2016 W23	219,853	162,420.2	223,956.7	225,149.8	28,575	26,663.13	28,992.93	29,798.77	71,678	65,404.02	71,854.38	74,317.41
2016 W24	209,304	162,420.2	206,502.8	218,590.6	28,739	26,663.13	29,664.07	29,918.09	68,981	65,404.02	75,802.32	76,751.86

B. Appendix B

B.1. R Code for Data Preprocessing

```
# Prepare the calendar data
calendar <- calendar_raw |>
  # Create a SNAP dummy column
  pivot_longer(
    cols = c(snap_CA, snap_TX, snap_WI),
    names_to = "state_id",
    values_to = "snap_dummy"
  ) |>
  # Get the last two characters to create a state id (CA, WI, TX)
  mutate(
    state_id = str_sub(state_id, -2, -1),
  ) |>
  arrange(date)
```

```
# Prepare the daily sales data
sales <- sales_raw |>
  pivot_longer(
    cols = -c(item_id, dept_id, cat_id, store_id, state_id),
    names_to = "day",
    values_to = "sales"
  ) |>
  filter(sales != 0)
```

```
# Merge sales with calendar and prices datasets
df <- sales |>
  # Join calendar data
```

```
left_join(  
  calendar,  
  by = join_by(day == d, state_id),  
  keep = FALSE  
) |>  
# Join prices data  
left_join(  
  prices,  
  by = join_by(wm_yr_wk, store_id, item_id),  
  keep = FALSE  
)
```

```
# Weekly aggregation of the sales dataset  
df_weekly <- df |>  
# Create a year-week column  
mutate(  
  week = yearweek(date, week_start = 1)  
) |>  
group_by(week, cat_id) |>  
summarise(  
  sales = sum(sales, na.rm = TRUE),  
  mean_price = mean(sell_price, na.rm = TRUE),  
  snap_days = n_distinct(  
    ifelse(snap_dummy == 1, date, NA), na.rm = TRUE  
  ),  
  event_days = n_distinct(  
    ifelse(!is.na(event_name_1), date, NA), na.rm = TRUE  
  )  
) |>  
ungroup() |>  
mutate(  
  month_lbl = month(week, label = TRUE, abbr = TRUE),  
  has_snap_over2 = as.factor(ifelse(snap_days > 2, 1, 0)),  
  has_event = as.factor(ifelse(event_days > 0, 1, 0))  
) |>  
# Get the week-of-month & mark last 2 weeks per month  
group_by(month, cat_id) |>
```

```
mutate(  
  month_week = row_number(),  
  is_last_2wk_mth = as.factor(  
    ifelse(month_week >= max(month_week) - 1, 1, 0)  
  )  
) |>  
ungroup() |>  
# Create lags of sales by category  
group_by(cat_id) |>  
arrange(week) |>  
mutate(  
  lag_1 = lag(sales, 1),  
  lag_2 = lag(sales, 2),  
  lag_3 = lag(sales, 3),  
  lag_4 = lag(sales, 4)  
)
```

B.2. R Code for Data Modeling

```
# Fit FOODS models on the training set  
fit_foods <- foods_train |>  
model(  
  # Linear Regression Model (LM)  
  lm_foods = TSLM(  
    sales ~ trend + month_lbl + is_last_2wk_mth +  
    has_snap_over2 + lag_1 + lag_4  
  ),  
  # Simple Exponential Smoothing (SES)  
  ses_foods = ETS(  
    sales ~ error("A") + trend("N") + season("N"),  
    alpha = 0.2032  
  ),  
  # Triple Exponential Smoothing (Holt-Winters)  
  hw_foods = ETS(  
    sales ~ trend("N") + season("N") + error("A"),  
    alpha = 0.2032, beta = 0.1, gamma = 0.1  
  )  
)
```

```
sales ~ error() + trend("A") + season("A"),  
alpha = 0.44,  
beta = 0.001,  
gamma = 0.92  
)  
)
```

```
# Fit HOBBIES models on the training set  
fit_hobbies <- hobbies_train |>  
model(  
  # Linear Regression Model (LM)  
  lm_hobbies = TSLM(  
    sales ~ trend + month_lbl + mean_price + lag_1 + lag_2 + lag_3  
  ),  
  # Simple Exponential Smoothing (SES)  
  ses_hobbies = ETS(  
    sales ~ error("A") + trend("N") + season("N"),  
    alpha = 0.3694  
  ),  
  # Triple Exponential Smoothing (Holt-Winters)  
  hw_hobbies = ETS(  
    sales ~ error("A") + trend("A") + season("A"),  
    alpha = 0.28,  
    beta = 0.01,  
    gamma = 0.77  
  )  
)
```

```
# Fit HOUSEHOLD models on the training set  
fit_hshld <- hshld_train |>  
model(  
  # Linear Regression Model (LM)  
  lm_hshld = TSLM(  
    sales ~ trend + month_lbl + is_last_2wk_mth +  
      has_snap_over2 + lag_1 + lag_2 + lag_3  
  ),  
  # Simple Exponential Smoothing (SES)
```

```
ses_hshld = ETS(  
  sales ~ error("A") + trend("N") + season("N"),  
  alpha = 0.5667  
)  
# Triple Exponential Smoothing (Holt-Winters)  
hw_hshld = ETS(  
  sales ~ error("A") + trend("A") + season("A"),  
  alpha = 0.11,  
  beta = 0.01,  
  gamma = 0.99  
)  
)
```

Author's Statement:

I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.