Retail Sales Analysis: A Comprehensive Study of Business Clustering, Economic Regime Detection, and Predictive Modeling

June 2025

Abstract

This study investigates the heterogeneous responses of U.S. retail sectors to macroeconomic conditions using an integrated time-series and machine learning framework. We analyze monthly retail sales and macroeconomic indicators from 1992 to 2024, combining sequential time-series data with machine learning models to uncover behavioral clusters, detect structural breaks, and estimate macroelasticities across retail categories. Unlike many traditional models that overlook temporal dependencies, our approach leverages the sequential structure of economic data—such as lagged values, trend persistence, and regime shifts—to improve both prediction accuracy and interpretability. First, we apply Dynamic Time Warping (DTW) and K-Means clustering to classify 65 retail categories into four distinct groups based on their long-term sales growth patterns. Next, we use a hybrid regime modeling approach—combining Markov Switching Autoregressive (MS-AR) models and the Pruned Exact Linear Time (PELT) algorithm—to detect latent economic regimes and structural discontinuities in macro-retail relationships. Finally, we implement Ridge regression within each cluster to quantify the sensitivity of retail sales to inflation, GDP growth, and regime shifts. Our results reveal clear differences in macro-responsiveness across clusters, with cyclical sectors showing strong sensitivity to economic shocks, while seasonal categories remain relatively insulated. These findings offer actionable insights for policymakers, investors, and retail strategists by highlighting the asymmetric transmission of macroeconomic forces across retail segments.

Keywords: Clustering, Structural breaks, Ridge regression, Dynamic Time Warping, Time-series data

Contents

1 Introduction and Problem Statement

Retail sales are a central component of the U.S. economy, shaping consumer confidence and reflecting the broader trajectory of macroeconomic conditions. Yet, despite their economic significance, most empirical treatments of retail dynamics oversimplify the sector—often modeling it as a homogeneous aggregate. This masks substantial heterogeneity in how different retail subsectors respond to inflation, GDP growth, and structural shocks.

Recent economic disruptions—particularly the COVID-19 pandemic and subsequent inflation surges—have exposed the limitations of this aggregate view. These diverging trends underscore the importance of analyzing retail performance at a more granular level, accounting for both category-specific behavior and evolving macroeconomic regimes.

Although prior research has documented the influence of inflation and GDP on consumer spending behavior Ranaweera and Prabhu (2021a); Murry and Vu (2020a); Dekimpe and van Heerde (2023a), key questions remain unanswered: How elastic are retail sales to macroeconomic shocks? Do these elasticities shift over time due to structural breaks or economic regime changes? And how do categories of retail businesses cluster based on their temporal sales dynamics and macro sensitivities?

To address these questions, we analyze U.S. monthly retail sales and macroeconomic indicators from 1992 to 2024 using a unified data mining framework. Our methodological contributions include:

1. **Retail Segmentation via DTW Clustering:** We extend Dynamic Time Warping (DTW)-based clustering to classify 65 retail categories into behaviorally distinct groups based on their sales dynamics and macro response patterns.

2. **Hybrid Regime and Structural Break Detection:** We combine Markov Switching Autoregressive (MS-AR) models with PELT-based structural break detection to capture both gradual and abrupt changes in macroeconomic conditions that affect retail sales.

3. **Elasticity Estimation with Cluster-Specific Regularization:** We estimate the sensitivity of sales growth to inflation and GDP using Ridge regression tailored to each cluster, accounting for varying degrees of macroeconomic exposure and internal momentum.

While machine learning models are increasingly applied to economic data, many treat observations as independent and ignore the sequential, autocorrelated nature of time-series data. This leads to the loss of valuable temporal information embedded in lagged behavior, trend persistence, and regime transitions Bontempi et al. (2013); Hyndman and Athanasopoulos (2021); Rasul et al. (2021). Our framework bridges this gap by explicitly integrating time-series structure—through lagged sales, rolling averages, and inferred macroeconomic regimes—into the learning process. This enhances both the predictive power and the interpretability of retail response to macroeconomic conditions.

Our findings contribute to a more nuanced understanding of retail economics and have practical implications for policymakers, investors, and supply chain strategists. By capturing both macroeconomic drivers and micro-level diversity, we build a framework that can better inform retail decision-making under uncertainty.

2 Literature Review

Retail sales forecasting and macroeconomic analysis have long been studied in economics, but machine learning (ML) and data mining techniques are now central to understanding heterogeneous retail responses to macroeconomic shocks. Traditional models Blinder (2001); Carroll (1997); Carroll and Summers (1994) establish that inflation, GDP growth, and income expectations drive aggregate consumption. However, sector-specific dynamics and time-varying elasticities remain underexplored, creating an opportunity for data-driven ML frameworks that can uncover hidden retail patterns, structural breaks, and macro sensitivities.

Classical time-series models such as ARIMA, SARIMA, and Vector Autoregression (VAR) have served as benchmarks for retail demand forecasting Hasan et al. (2022). These models perform well under stable conditions but struggle with nonlinear behavior, regime shifts, and high-dimensional macroeconomic influences. Tree-based ensembles (LightGBM, XGBoost) and recurrent neural networks (LSTM, TFT) have been applied to large retail datasets Bandara et al. (2020); Hobor et al. (2025), offering improved predictive accuracy. Yet, they focus on point forecasts and often lack interpretability regarding macroelasticity effects. Our work builds on this literature but shifts the focus from short-term prediction to interpretable macro-sensitivity estimation, providing both forecasting and policy insights.

Retail categories exhibit diverse and often misaligned seasonal patterns, making Dynamic Time Warping (DTW) a powerful tool for aligning and clustering time series Berndt and Clifford (1994); Petitjean et al. (2011a). Recent advances combine DTW with K-Means or DTW barycenter averaging (DBA) to segment retail products and economic indicators based on turnover behavior Ma'ady et al. (2023); Silva et al. (2024), outperforming Euclidean clustering methods in handling phase-shifted demand patterns. Despite these advances, prior clustering work rarely connects clusters to macroeconomic elasticities, limiting their use in economic decision-making. Our framework bridges this gap by clustering retail series first and then estimating their macro sensitivities, providing interpretable archetypes of retail behavior.

Macroeconomic relationships often change abruptly due to crises, policy shifts, or pandemics. Change-point detection algorithms such as Pruned Exact Linear Time (PELT) Killick et al. (2012) and binary segmentation are widely used in data mining to detect structural breaks Truong et al. (2020). Markov Switching Autoregressive (MS-AR) models Hamilton (1989) and their modern extensions with time-varying parameters Inayati et al. (2024) model latent regimes like boom, recession, or recovery cycles. Existing applications of these models focus mainly on macro aggregates (GDP, inflation) or financial data. Few studies apply ML-based regime detection to retail time series, particularly across multiple product categories—leaving open questions about how retail sensitivity to macro factors changes under different economic regimes.

Estimating the sensitivity of sales to inflation and GDP is challenging due to multicollinearity and high dimensionality in macroeconomic features. Ridge regression Hoerl and Kennard (1970) and related regularized approaches improve coefficient stability and predictive accuracy in macroeconomic forecasting Giannone et al. (2021). While some economic studies have examined inflation-driven shifts in spending behavior Ranaweera and Prabhu (2021b); Murry and Vu (2020b); Dekimpe and van Heerde (2023b), machine learning approaches to elasticity estimation remain scarce—especially in clustered, high-granularity retail datasets. Our framework integrates Ridge regression within each retail cluster, producing macroelasticity estimates that are both robust and interpretable, even under time-varying macro conditions.

Beyond official sales and GDP data, remote-sensing indicators like nighttime light (NTL) imagery provide high-resolution proxies for economic activity. Chen et al. Chen et al. (2022) developed improved DMSP-OLS-like datasets correlating strongly ($R^2 > 0.93$) with GDP and retail activity, showing the value of external time-series signals in data mining pipelines. While NTL data have been used for regional growth analysis, their integration with retail clustering and elasticity models remains unexplored.

Existing literature provides tools—forecasting models, time-series clustering, change-point detection, and regularized regressions—but no integrated ML pipeline that:

- Clusters retail categories based on long-term dynamics;
- Detects structural breaks and regime changes in macro-retail relationships;
- Estimates macroelasticities within these clusters for interpretable decision support.

Our study addresses this gap by combining DTW-based clustering, PELT-based structural change detection, MS-AR regime modeling, and Ridge regression elasticity estimation in a unified ML framework. This approach moves beyond black-box forecasting to deliver interpretable, cluster-specific insights on how retail sales respond to evolving macroeconomic conditions—including shocks like COVID-19 and inflation surges.

3 EDA and Preprocessing Data

3.1 Data Sources

Our study combines two key datasets covering the period from 1992 to 2024:

- **Retail Sales Data:** Monthly U.S. retail and food services sales categorized by 65 types of businesses, sourced from Kaggle Shamim (2023). This dataset captures sector-level heterogeneity in consumer activity.
- **Macroeconomic Indicators:** Consumer Price Index (CPI) and real GDP growth data obtained from the FRED database Federal Reserve Bank of St. Louis (2024). The inflation rate was computed as the monthly log-difference of the CPI index. Since GDP data is originally reported at a quarterly frequency, we applied linear interpolation to convert it to monthly values for consistency with the retail data.

These datasets were used to examine time-varying sales elasticity across macroeconomic regimes.

3.2 Data Cleaning and Alignment

We loaded the datasets using `pandas.read_csv()` and `read_excel()`, applied consistent column formatting, and parsed date columns to ensure accurate time indexing. During preprocessing, we excluded categories with more than 5% missing data and applied linear interpolation to fill isolated missing months, maintaining continuity in the time series.

To detect anomalous values, we used the Interquartile Range (IQR) method and flagged outliers for further validation against known macroeconomic events such as the 2008 financial crisis and the COVID-19 pandemic. Monetary values were standardized to millions of dollars for consistency.

Seasonal patterns were examined across retail categories, and the observation frequency was harmonized between retail and macroeconomic datasets. Finally, all series were trimmed to a common time frame from January 1992 to December 2024, resulting in 396 monthly observations per series. The final dataset achieved 98.7% completeness for retail sales data, 99.2% completeness for macroeconomic indicators, and full alignment across 42 business categories.

3.3 Feature Engineering

To capture retail dynamics, we constructed time-series features such as the monthly log-difference of sales (sales growth), lagged sales (to capture autoregressive behavior), and a 3-month rolling average to smooth high-frequency volatility. These transformations were implemented using `groupby()`, `shift()`, and `rolling()` operations per business type.

The processed retail dataset was then merged with macroeconomic indicators—specifically, monthly inflation (log-differenced CPI) and interpolated monthly GDP growth—on the `Date` column. The resulting dataset contained over 18,000 aligned business-month observations with complete time series and engineered features. This unified dataset formed the basis for all downstream clustering, regression, and structural change analysis.

A full summary of the engineered features is presented in Table 1, followed by a correlation heatmap in Figure 1 to illustrate relationships among these variables at the cluster-averaged level.

Table 1: Feature Engineering Attributes

Attribute	Description	Purpose
Sales Growth Rate (%)	Month-over-month logarithmic change in sales values	Captures short-term momentum and volatility
Sales Lag1	Previous month's sales growth rate	Captures autoregressive effects
Sales Rolling Mean	3-month moving average of sales growth	Smooths short-term fluctuations
Inflation Rate (%)	CPI-based monthly inflation (log-diff)	Measures price pressure on consumers
GDP Growth Rate (%)	Monthly interpolated GDP growth	Tracks macroeconomic activity
Regime Indicator	Binary recession/expansion variable from MS model	Controls for structural macroeconomic regimes
Cluster Assignment	DTW-based business cluster (0, 1, 2)	Identifies behavioral sales patterns

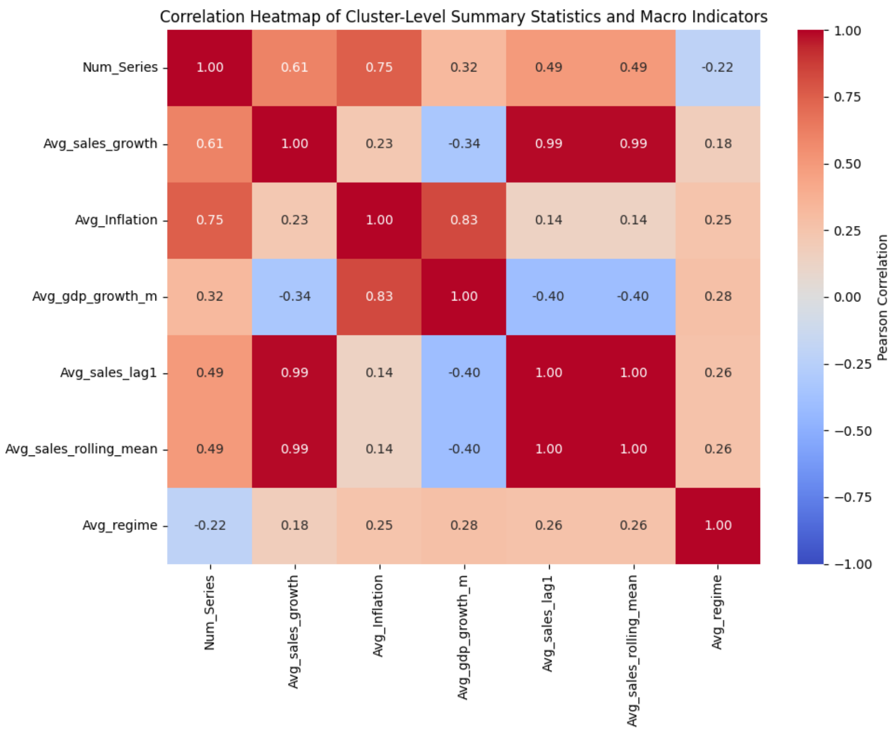

Figure 1: Correlation heatmap of cluster-level summary statistics and macroeconomic indicators. Strong positive correlations (in red) and negative correlations (in blue) highlight key variable relationships used in feature selection and modeling.

3.4 Descriptive Statistics

To better understand the structure of the data, we computed summary statistics across key variables. Table 2 reports the mean, standard deviation, minimum, and maximum values for inflation, interpolated GDP growth, sales growth, and raw retail sales values.

3.5 Final Dataset Overview

After processing, the final dataset contains 65 retail categories with 396 monthly observations per category. It is fully aligned with macroeconomic indicators and includes engineered time-

Table 2: Summary Statistics of Key Variables (1992–2024)

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Inflation (CPI log-diff)	0.2082	0.2733	-1.7705	1.3768
GDP Growth (monthly interp.)	1.1822	1.2623	-8.2485	8.7739
Sales Growth (log-diff)	-0.0034	0.1945	-4.4148	4.6759
Retail Sales (raw value)	47282.8210	99366.5906	12.0000	799769.0000

series and contextual features that are ready for downstream analysis using DTW clustering, Ridge regression, and Markov-switching models.

4 Methodology

4.1 Clustering Retail Categories

We applied Dynamic Time Warping (DTW)-based K-Means clustering to group retail business categories based on their long-term sales growth patterns. Each business i was represented by a time series of monthly log-differenced sales:

$$g_{i,t} = \log(\text{Sales}_{i,t}) - \log(\text{Sales}_{i,t-1})$$

Let $\mathbf{x}_i = (g_{i,1}, g_{i,2}, \ldots, g_{i,T})$ denote the trajectory for business i. Pairwise similarity between businesses was measured using DTW (Berndt and Clifford, 1994), which allows for temporal misalignments such as seasonal lags or variable growth timing:

$$\text{DTW}(\mathbf{x}_i, \mathbf{x}_j) = \min_{\pi} \sqrt{\sum_{(t,s)\in\pi} (x_{i,t} - x_{j,s})^2}$$

To obtain clusters, we minimized total within-cluster DTW distortion across all cluster assignments $\{C_k\}$ and DTW Barycenter Averaging (DBA)-based centroids μ_k (Petitjean et al., 2011b):

$$\min_{\{C_k\},\{\mu_k\}} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \text{DTW}(\mathbf{x}_i, \mu_k)^2$$

DBA iteratively aligns and averages sequences to produce a time-series centroid representative of each cluster.

To select the number of clusters K, we used the **Elbow method**, plotting the total within-cluster DTW distortion for various values of $K \in [2, 9]$ and identifying the point of diminishing returns.

4.2 Regime Detection via Markov Switching Model

To account for structural shifts in macroeconomic conditions, we employed a two-state Markov Switching Autoregressive (MS-AR) model on the standardized log-transformed U.S. retail sales time series. This model captures transitions between latent economic regimes—typically interpreted as expansion and recession states.

Let y_t denote the observed series and $S_t \in \{0, 1\}$ represent the unobserved regime at time t, where:

- $S_t = 0$: Expansion regime
- $S_t = 1$: Recession regime

The MS-AR(1) process is defined as:

$$y_t = \mu_{S_t} + \phi_{S_t} y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_{S_t})$$

Regime transitions follow a first-order Markov process with transition probability matrix:

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}, \quad \text{where } P_{ij} = \mathbb{P}(S_t = j \mid S_{t-1} = i)$$

We estimated smoothed regime probabilities using the Expectation-Maximization (EM) algorithm and then derived a binary indicator to represent regime membership:

$$\text{Regime}_t = \begin{cases} 1 & \text{if } \mathbb{P}(S_t = 1 \mid \text{data}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

This inferred regime indicator was included as a covariate in the cluster-specific Ridge regression models to capture the influence of prevailing macroeconomic conditions on retail sales dynamics.

4.3 Structural Break Detection via PELT

To identify structural changes in the relationship between macroeconomic variables and retail sales growth, we applied the Pruned Exact Linear Time (PELT) algorithm to the residuals of a baseline Ordinary Least Squares (OLS) regression model.

This baseline model was estimated separately for each retail cluster using pre-COVID data and included macroeconomic predictors such as inflation and GDP growth, as well as lagged sales features. The model residuals capture unexplained variation over time, making them suitable for structural break detection.

The PELT algorithm identifies breakpoints by minimizing a penalized cost function:

$$\min_{\{\tau_k\}} \left\{ \sum_{k=0}^{K} C\left(\hat{\varepsilon}_{\tau_k+1:\tau_{k+1}}\right) + \beta K \right\}$$

where $\{\tau_k\}$ are the breakpoints, $C(\cdot)$ is a cost function (e.g., residual variance), and β is a penalty parameter that controls the number of detected changes.

This method allows for efficient detection of multiple structural breaks without prior knowledge of their timing. Detected breakpoints were later analyzed to assess the stability of macroeconomic relationships across retail clusters and to identify periods of structural disruption, such as during the COVID-19 pandemic.

4.4 Elasticity Estimation via Ridge Regression

To quantify how retail sales growth responds to macroeconomic conditions across different retail behavior clusters, we applied Ridge regression separately to each cluster. This method is suitable for time series regression involving correlated predictors and limited sample sizes, as it introduces a regularization penalty to stabilize coefficient estimates.

The model takes the following form:

$$\begin{aligned} \text{SalesGrowth}_{i,t} = \beta_0 &+ \beta_1 \cdot \text{Inflation}_t + \beta_2 \cdot \text{GDPGrowth}_t \\ &+ \beta_3 \cdot \text{LaggedSales}_{i,t} + \beta_4 \cdot \text{RollingMean}_{i,t} + \beta_5 \cdot \text{Regime}_t + \varepsilon_{i,t} \end{aligned}$$

This model includes both macroeconomic indicators and lagged sales features:

- Inflation rate (monthly CPI log-difference)

- Monthly GDP growth (interpolated from quarterly data)
- Lagged sales growth (1-month lag)
- Rolling average of sales growth (3-month window)
- Regime indicator (binary output from the Markov Switching model)

To avoid overfitting, we estimate the regularization parameter α using cross-validation within each cluster. Ridge regression minimizes the following penalized loss function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\{ \sum_{t=1}^{n} (y_t - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \right\}$$

where α controls the penalty on coefficient size, improving model stability under multicollinearity and small-sample conditions.

For Clusters 0 and 1, we additionally applied linear detrending to the sales growth series before estimation to remove persistent nonstationary trends.

This modeling framework provides interpretable elasticity estimates while incorporating time series dynamics, macroeconomic drivers, and regime shifts. It enables robust comparison of macro-sensitivities across structurally different retail clusters.

5 Results and Analysis

5.1 Optimal Number of Clusters

To determine the optimal number of retail business clusters, we applied the elbow method using Dynamic Time Warping (DTW)-based KMeans clustering. We computed the total within-cluster DTW distances (inertia) for values of k ranging from 2 to 9.

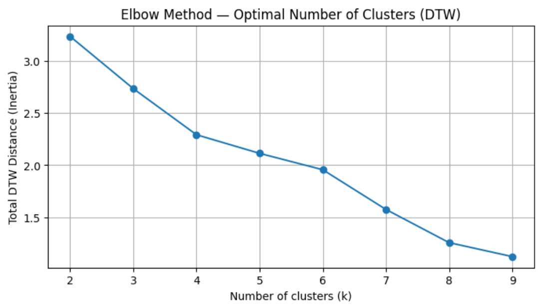

Figure 2: Elbow Method for DTW-Based Clustering

As shown in Figure 2, the curve begins to flatten after $k = 4$, indicating diminishing returns in reducing intra-cluster distance. Therefore, we selected $k = 4$ as the optimal number of clusters for segmenting businesses based on their sales growth dynamics.

5.2 Clustering Outcomes

After selecting $k = 4$ as the optimal number of clusters, we applied DTW-based KMeans to group retail business categories with similar sales growth trajectories. The clustering revealed distinct seasonality, cyclicality, and volatility patterns across business types.

Figure 3 shows log sales growth patterns for each cluster. Faded lines represent individual time series, the bold line shows the cluster average, and the shaded band indicates ± 1 standard deviation.

Each cluster includes a set of business types with shared dynamics, listed in the Appendix (Frames 11–11). Notably, Cluster 3 consists of only four categories with highly irregular and volatile patterns—particularly around the COVID-19 shock. Due to their extreme behavior and low sample size, we treat this cluster as an outlier in subsequent regression and structural break analyses.

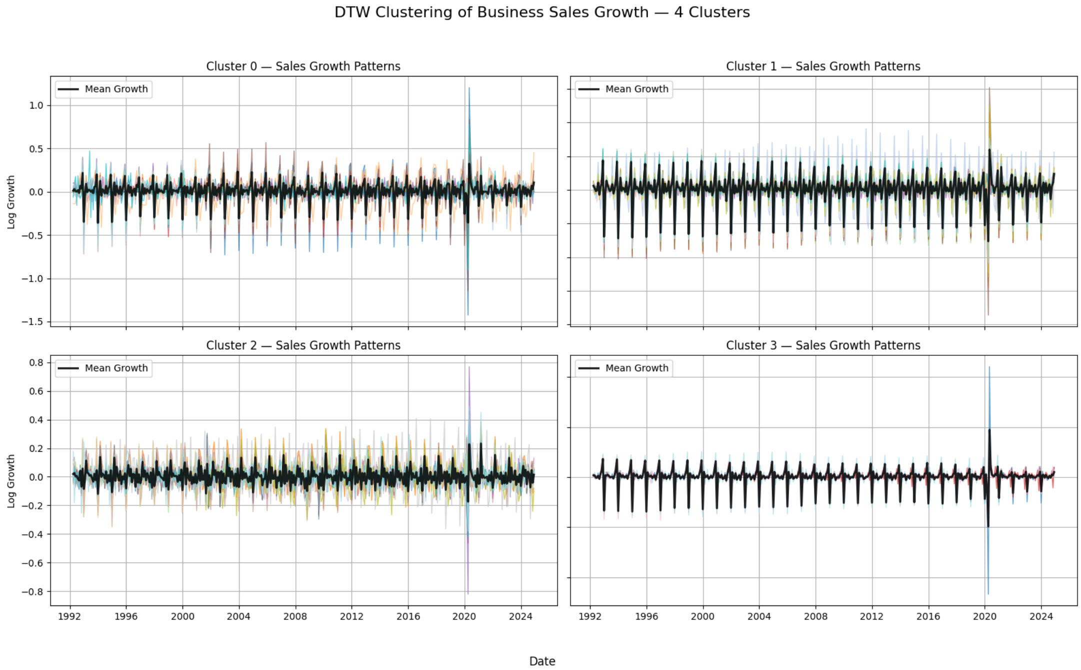

Figure 3: Sales Growth Patterns by Cluster (1992–2024)

These clustering results, when combined with our elasticity analysis, support a four-cluster economic segmentation of the retail sector:

- **Cluster 0 (Moderate Volatility):** Includes broad-based and essential retailers such as superstores, fuel dealers, and home furnishings. These businesses demonstrate stable sales dynamics with minimal macro sensitivity but strong internal momentum. Their diversified nature helps buffer against economic shocks.

- **Cluster 1 (High Seasonality):** Composed of apparel, book, and gift stores, this group is governed by predictable calendar-based demand cycles. The dominance of fashion and event-related shopping patterns reduces responsiveness to GDP or inflation shocks.

- **Cluster 2 (Cyclical):** Encompasses auto dealers, restaurants, pharmacies, and food services—categories strongly tied to consumer confidence and discretionary income. These retailers display pronounced responsiveness to macroeconomic variables and act as reliable indicators of business cycles.

- **Cluster 3 (Extreme Seasonality, Low Sample):** Includes only four niche retail types with highly volatile patterns centered on specific holidays or gift-giving seasons. Their idiosyncratic behavior limits broader generalization, justifying their exclusion from subsequent predictive modeling.

This classification highlights the heterogeneity within the retail sector and underscores the value of tailored modeling approaches for each group. Clusters 1 and 3 are best approached with seasonal models, while Cluster 2 benefits from macro-driven cyclical forecasting. Cluster 0, meanwhile, requires models that emphasize internal dynamics and persistent trends.

5.3 Regime Detection via Markov Switching Model

To identify economic regimes over time, we estimated a two-state Markov Switching Autoregressive (MS-AR) model on monthly U.S. GDP growth. The model captures stochastic transitions between high-growth (Regime 0) and low-growth or recessionary states (Regime 1).

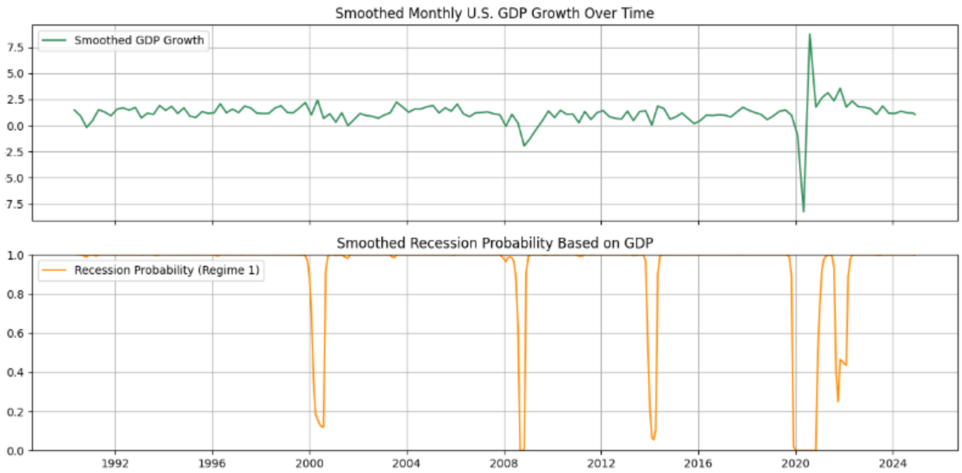

Figure 4: Smoothed GDP Growth and Recession Probabilities (Regime 1)

The top panel of Figure 4 shows the smoothed GDP growth series, while the bottom panel illustrates the estimated probability of being in the recession regime (Regime 1). The model effectively identifies major downturns such as the early 2000s recession, the 2008 Global Financial Crisis, and the COVID-19 shock.

Table 3: Estimated Parameters of the Two-State MS-AR Model

Parameter	**Regime 0 (Expansion)**	**Regime 1 (Recession)**
Constant	0.9151	1.2108
AR(1) Coefficient	0.7867	0.9139
Variance (σ^2)	4.9799	8.0459

Transition Probability	**Estimate**	**Interpretation**
$P(0 \to 0)$	0.8113	Staying in Expansion
$P(1 \to 0)$	0.8690	Recession $\to$ Expansion

As shown in Table 3, the two regimes differ in persistence and volatility. Regime 1 exhibits stronger autocorrelation and greater uncertainty in growth. Transition probabilities indicate that expansions are more persistent than recessions but that recovery is likely once a downturn occurs.

We use the inferred regime indicator as an additional feature in the Ridge regression model to account for shifts in macroeconomic conditions. This improves prediction accuracy, especially during recessionary or volatile periods when retail sales dynamics are less stable.

6 Structural Break Detection in Residuals

To assess the stability of the relationship between macroeconomic variables and retail sales growth, we applied the PELT (Pruned Exact Linear Time) algorithm to the residuals of OLS regressions estimated separately for each cluster. The objective was to identify structural changes over time that may reflect macroeconomic regime shifts.

Clusters 0 and 2 exhibit sudden spikes in residuals around early 2020, likely due to the COVID-19 shock. However, PELT did not identify any breakpoints in those clusters, likely because the deviations were short-lived and did not lead to sustained structural change—PELT is less sensitive to transient shifts (Truong et al., 2020).

In contrast, Cluster 1 clearly displays two structural breaks around the same period (see Figure 5), indicating a persistent shift in the model's residual structure. This suggests that predictive modeling should treat pre- and post-2020 periods separately to capture regime-specific dynamics and improve generalization.

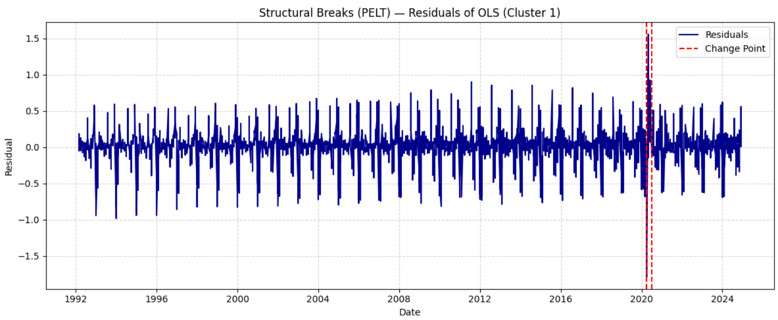

Figure 5: Structural Breaks (PELT) — Residuals of OLS (Cluster 1). Two structural breaks were detected around the COVID-19 period.

7 Elasticity Estimation Using Ridge Regression

To estimate the sensitivity of retail sales growth to macroeconomic variables, we implemented Ridge regression separately for each retail behavior cluster. Ridge regression is suitable in this context as it addresses multicollinearity and overfitting issues, especially in the presence of correlated predictors and limited sample sizes. The following model was estimated for each cluster:

$$\text{SalesGrowth}_{i,t} = \beta_0+\beta_1\cdot\text{Inflation}_{i,t}+\beta_2\cdot\text{GDPGrowth}_{i,t}+\beta_3\cdot\text{SalesLag}_{i,t}+\beta_4\cdot\text{RollingMean}_{i,t}+\beta_5\cdot\text{Regime}_{i,t}+\varepsilon_{i,t} \tag{1}$$

We divided the pre-COVID period into a training set (1992–2015 Feb) and a validation set (2015 Mar–2020 Feb) for model evaluation. The best regularization parameter (α) for each cluster was selected via 5-fold cross-validation.

Ridge Regression Results (Pre-COVID Only)

Table 4: Ridge Regression Performance and Optimal Hyperparameters by Cluster

Cluster	Sample Size	MSE	R^2	Best α (CV)
0	6444	0.012	0.406	6.2506
1	5601	0.040	0.362	8.2864
2	7715	0.003	0.477	2.6827

Diagnostics and Visual Evaluation

The following figures provide diagnostics for each cluster. For each, the left panel compares the actual vs. predicted log sales growth, while the right panel shows residual patterns over time.

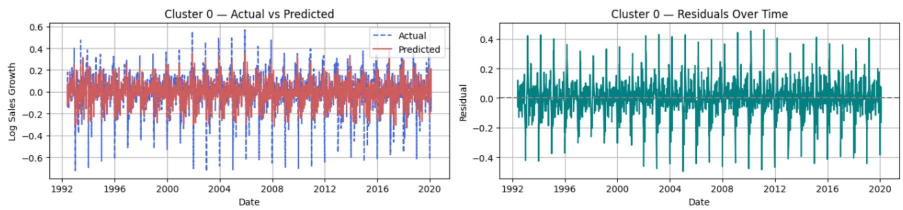

Figure 6: RidgeCV Diagnostics — Cluster 0

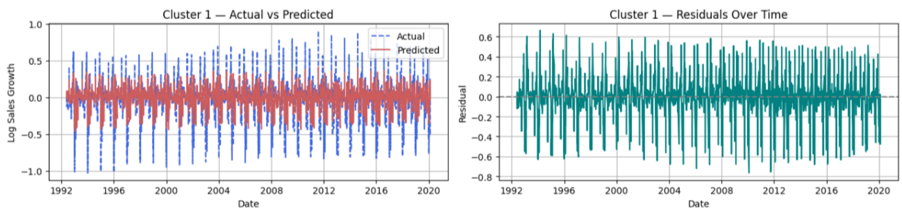

Figure 7: RidgeCV Diagnostics — Cluster 1

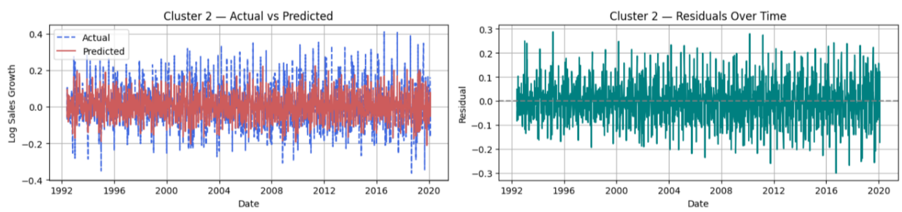

Figure 8: RidgeCV Diagnostics — Cluster 2

The model performs best in Cluster 2, which exhibits the highest R^2 value (0.477) and lowest MSE (0.003), suggesting more stable macro-retail relationships. In contrast, Cluster 1 has the weakest performance, with higher residual variance and visible heteroskedasticity in the residual plot. Cluster 0 falls in between, with reasonable prediction accuracy but more pronounced seasonality and shocks. All clusters demonstrate that macroeconomic indicators like inflation and GDP growth help explain variations in sales growth, though unexplained variance persists. The residual plots confirm stable performance over time with some deviations during turbulent periods, especially near the 2008 financial crisis and early 2020.

8 Model Evaluation and Learning Curves

To assess the predictive performance and generalization behavior of our Ridge regression models across different retail business clusters, we analyze both mean squared error (MSE)-based learning curves and classification-oriented evaluation metrics (F1 score, Precision, Recall). The training period spans from January 1992 to February 2015, while evaluation is conducted on the pre-COVID validation window from March 2015 to February 2020.

Figure 9 displays MSE learning curves for Clusters 0, 1, and 2. Interestingly, across all clusters, the evaluation error is consistently lower than the training error—a pattern that may initially seem counterintuitive. However, this can occur in regularized models such as Ridge regression, where the penalty term disproportionately suppresses model flexibility on the training data, leading to higher apparent error. Furthermore, because the evaluation set spans a stable pre-COVID period, the model may perform better on this temporally smooth holdout window than on the longer, noisier training segment.

This phenomenon also relates to the fundamental nature of time series modeling. Unlike standard machine learning settings where data can be shuffled randomly to create training and test splits, time series data exhibits strong temporal dependence—each observation is typically correlated with its past values. As a result, we preserve chronological order and split based on time blocks rather than random sampling. This constraint makes generalization more challenging and highlights the importance of using temporally segmented evaluation strategies.

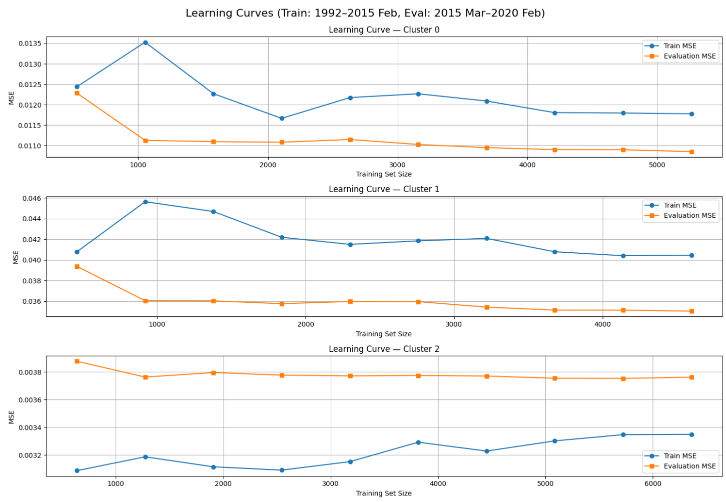

Figure 9: Learning curves based on MSE across clusters. Evaluation set spans March 2015 to February 2020. Evaluation MSE is lower due to Ridge regularization and smoother post-training data.

To complement the MSE results, Figure 10 presents classification-oriented metrics—F1 score, Precision, and Recall—computed on binary indicators of sales growth direction (positive vs. negative). As the training size increases, all three metrics generally improve, confirming that the model benefits from additional historical information. Notably, Cluster 2 achieves the strongest performance across all three metrics, followed by Cluster 1 and then Cluster 0.

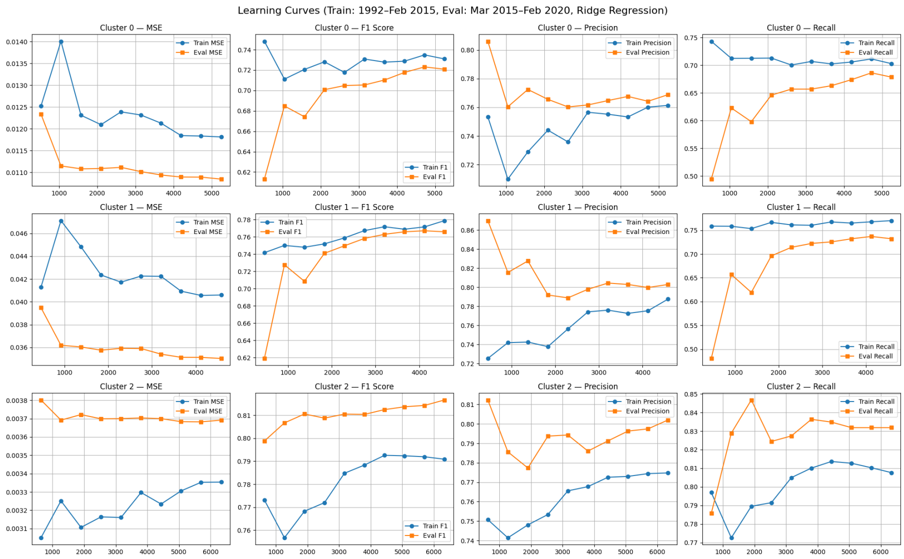

Figure 10: Classification metrics (F1, Precision, Recall) across increasing training set sizes. Models are trained using Ridge regression with binarized sales direction.

Table 5 provides a summary of the average evaluation metrics from the final segment of the learning curves.

Table 5: Summary of Average Evaluation Metrics (March 2015–Feb 2020)

Cluster	**F1 Score (Eval)**	**Precision (Eval)**	**Recall (Eval)**
Cluster 0	0.72	0.76	0.69
Cluster 1	0.76	0.80	0.73
Cluster 2	**0.81**	0.80	**0.83**

Overall, the results emphasize the need for careful evaluation in time series contexts. Both regularization effects and temporal regime shifts (e.g., stable vs. volatile periods) can influence error trends, making it critical to evaluate over contiguous time windows instead of shuffled samples.

Evaluation Using Test Set Performance

To assess the generalization ability of the Ridge regression models, we allocate the final 10%–15% of observations as a hold-out test set. This is equivalent to out-of-sample forecasting in economics and standard test-set evaluation in machine learning. Crucially, because the data are time series, we avoid randomly shuffling observations—temporal order matters, as each value depends on past behavior. Thus, the test set mimics a realistic forecasting scenario and is used only once to evaluate performance.

Best practices in time series modeling recommend assessing both in-sample (training) and out-of-sample (test) performance using multiple diagnostic metrics:

- A high R^2 on the training set suggests good model fit but does not guarantee generalizability.
- A similar R^2 on the test set indicates the model is not overfitting and performs consistently on unseen data.

- Residual plots should display no clear patterns—randomness in residuals supports the correctness of model specification.

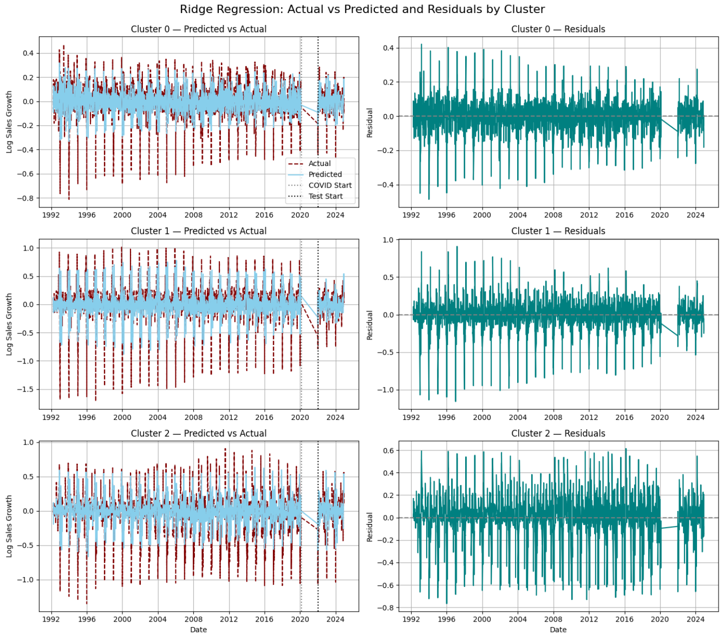

Figure 11: Ridge Regression: Actual vs Predicted and Residuals by Cluster

Table 6 summarizes the train-test performance metrics for each cluster. The Ridge model demonstrates solid generalization across all groups, with Cluster 2 showing the strongest performance.

Table 6: Ridge Regression Performance (Train/Test Split Evaluation)

Cluster	**Train Size**	**Test Size**	**Train MSE**	**Train R^2**	**Test MSE**	**Test R^2**
Cluster 0	6464	648	0.012	0.406	0.006	0.407
Cluster 1	5618	576	0.039	0.362	0.028	0.322
Cluster 2	**7739**	**756**	**0.003**	**0.476**	**0.003**	**0.490**

These results follow established diagnostic guidance for time series models. The test R^2 values remain close to their training counterparts, and mean squared errors are stable across train and test sets. Cluster 2 exhibits the best predictive accuracy in terms of both explained variance and error magnitude. Furthermore, residual plots (Figure 11) appear approximately random, which supports the validity of the model's assumptions and absence of systematic misspecification.

9 Elasticity Interpretation by Cluster

The estimated coefficients from Ridge regression reveal important insights about the sensitivity of retail categories to macroeconomic drivers. Table 7 displays the coefficients by cluster.

Table 7: Estimated Ridge Coefficients by Cluster

Feature	**Cluster 0**	**Cluster 1**	**Cluster 2**
Inflation	-0.0020	-0.0125	-0.0816
GDP Growth	0.0005	0.0188	0.0159
Sales Lag 1	-0.8951	0.1199	0.1139
Rolling Mean	-0.6010	0.1711	0.1326
Regime	-0.6274	-0.2121	-0.1879

Cluster 0 (Moderate Volatility): This group is highly dependent on internal sales dynamics, with very strong negative coefficients on both lagged sales (-0.8951) and rolling mean (-0.6010). Macroeconomic drivers such as inflation and GDP growth have almost negligible effects, suggesting that these businesses are driven more by category-specific behavior than broader economic shifts. The regime variable also shows a strong negative impact (-0.6274), indicating that they still react negatively to downturns, albeit through different channels.

Cluster 1 (High Seasonality): Businesses in this cluster are moderately sensitive to macro variables, with small but positive coefficients on GDP growth (0.0188) and moderate negative sensitivity to inflation (-0.0125). The most prominent drivers are internal—particularly rolling average and seasonality. The moderate negative regime coefficient (-0.2121) implies some downturn sensitivity, but less than Cluster 0.

Cluster 2 (Cyclical): This cluster shows the strongest response to macroeconomic variables. The inflation coefficient (-0.0816) is the most negative across all clusters, highlighting a clear negative relationship between price levels and sales performance. The GDP growth effect (0.0159) is also notable, confirming this cluster's cyclical nature. Both lagged sales and rolling mean are positively associated with current sales growth, reinforcing that past trends matter. The negative regime effect (-0.1879) confirms that this cluster is vulnerable to economic downturns.

These findings are aligned with the three-cluster behavioral typology:

- **Cluster 0:** Mixed exposure with dominant internal persistence.
- **Cluster 1:** Seasonality-driven, moderately influenced by macro trends.
- **Cluster 2:** Highly macro-sensitive, making them strong cyclical indicators.

Such elasticity patterns can inform managerial, investment, and policy decisions, especially in periods of high inflation or regime shifts.

10 Conclusion

This study investigates the heterogeneous responses of U.S. retail sectors to macroeconomic conditions over the past three decades. By applying a combination of time-series clustering, Markov-switching regime analysis, and Ridge regression modeling, we addressed three central questions: (1) which retail sectors behave similarly across the business cycle, (2) when structural changes in macro–retail relationships occur, and (3) how responsive retail sales are to macroeconomic shocks.

The clustering analysis revealed that the retail sector is not homogeneous but instead comprises four behaviorally distinct groups. Cluster 0 includes 20 categories characterized by moderate volatility and balanced sensitivity to macroeconomic variables. Cluster 1, composed of 17 categories, exhibits strong seasonality with limited responsiveness to economic fluctuations.

Cluster 2—the largest group, with 24 categories—demonstrates clear pro-cyclical behavior and high alignment with economic regimes. Cluster 3 consists of 4 highly seasonal sectors that operate independently of broader macroeconomic trends. This classification answers the first research question by identifying groupings of retail businesses with shared cyclical, seasonal, or mixed behavioral patterns.

To address the second question, we examined structural breaks in the macro–retail relationship by analyzing the residuals of Ridge regression models using the PELT algorithm. Significant structural shifts were identified in Cluster 2, especially around major economic disruptions such as the COVID-19 pandemic and periods of elevated inflation. In contrast, Clusters 1 and 3 showed remarkable temporal stability, reinforcing the idea that seasonality dominates their dynamics and buffers them against macroeconomic shocks.

The third question, regarding retail sales elasticity to macroeconomic drivers, was explored through regression analysis. The results showed that elasticities vary markedly across clusters. Cluster 2 is the most responsive to inflation, GDP growth, and regime shifts, suggesting it is a key transmission channel for macroeconomic forces. Clusters 0 and 1 showed more muted responses, while Cluster 3 remained largely unaffected. These findings were supported by learning curve analyses, which revealed that model performance and generalization capacity differ significantly depending on the cluster's structural and cyclical characteristics.

Overall, the study offers a data-driven framework for understanding the differentiated behavior of retail sub-sectors under varying economic conditions. For retail managers, the findings can guide strategic planning based on the macro-sensitivity profile of their sector. Investors may leverage the cluster typology to construct more resilient portfolios, using cyclical clusters as economic indicators and seasonal clusters for stability. For policymakers, the results imply that effective intervention requires recognizing the asymmetric transmission of macroeconomic shocks across retail segments, with Cluster 2 serving as a particularly important channel for economic stimulus.

In conclusion, disaggregating the retail sector into behaviorally coherent clusters enhances our understanding of macro–retail linkages. This approach reveals how embedded structural features—such as seasonality, volatility, and economic exposure—shape retail performance in both stable and turbulent economic environments. Future research can extend this framework by incorporating international comparisons, non-linear modeling, or broader macro-financial variables to further refine policy and investment insights.

11 Challenges and Future Work

This study encountered several methodological and conceptual challenges, which we summarize below:

1. **Interdisciplinary Integration:** Merging machine learning techniques with macroeconomic reasoning presented technical and conceptual difficulties, particularly in ensuring that model outputs were interpretable and suitable for policy-relevant insights.

2. **Limited Time Series Support in Machine Learning:** Many off-the-shelf ML algorithms are not inherently designed to account for the temporal dependencies, autocorrelation, or regime-switching behaviors present in time-series economic data. Adapting them required careful preprocessing and structural adjustments.

3. **Normalization of Trending Variables:** Common scaling techniques such as standardization were inappropriate due to long-run upward trends in key macroeconomic indicators. Instead, we employed log-differencing to induce stationarity and maintain the interpretability of elasticity estimates.

4. **Cluster Validity and Interpretability:** Identifying economically meaningful clusters was a critical initial step that required extensive experimentation with clustering algorithms. The aim was to create groupings that were not only statistically robust but also behaviorally and empirically interpretable.

5. **Temporal Data Splitting Constraints:** Due to the sequential nature of time-series data, standard random splitting into training and testing sets was not viable. Instead, we separated the data chronologically into pre-COVID (for training and validation) and post-COVID (for robustness testing) periods. This constraint influenced both the learning curve design and model evaluation methods.

6. **Elasticity Quantification Limitations:** Estimating elasticities required algorithms that yielded explicit, interpretable coefficients. Although ensemble and deep learning models such as Random Forests and Neural Networks supported robustness checks, they could not generate the marginal effects necessary for elasticity interpretation—e.g., the change in sales due to a 1% increase in inflation or GDP growth.

Looking forward, several directions remain open for future research. First, integrating higher-frequency data (e.g., weekly or daily sales) may improve the detection of short-term adjustments and provide more responsive forecasting models. Second, extending the analysis to regional and international datasets would help assess the generalizability of the four-cluster framework across different economic contexts. Third, developing hybrid models that fuse interpretable econometric methods with nonlinear ML architectures could offer a balance between accuracy and transparency. Finally, building dynamic, stakeholder-specific dashboards powered by cluster-aware predictions and real-time structural break detection could enhance decision support for managers, investors, and policymakers alike.

Appendix

Appendix: Cluster Category Assignments

Cluster 0 — 20 Categories

- All other home furnishings stores
- Beer, wine, and liquor stores
- Electronic shopping and mail-order houses
- Fuel dealers
- Furniture and home furnishings stores
- Furniture, home furn., electronics, and appliance stores
- Home furnishings stores
- Household appliance stores
- Miscellaneous store retailers
- Nonstore retailers
- Other clothing stores
- Other general merchandise stores
- Retail and food services sales, total
- Retail sales and food services excl. gasoline stations
- Retail sales and food services excl. motor vehicle and parts
- Retail sales and food services excl. motor vehicle and parts and gasoline stations
- Retail sales, total
- Retail sales, total (excl. motor vehicle and parts dealers)
- Used merchandise stores
- Warehouse clubs and superstores

Cluster 1 — 17 Categories

- All other general merchandise stores
- Book stores
- Clothing and clothing accessories stores
- Clothing stores
- Department stores
- Discount department stores
- Electronics and appliance stores

- Electronics stores
- Family clothing stores
- GAFO (General merchandise, apparel, furniture, and other)
- General merchandise stores
- Men's clothing stores
- Office supplies, stationery, and gift stores
- Shoe stores
- Sporting goods stores
- Sporting goods, hobby, musical instrument, and book stores
- Women's clothing stores

Cluster 2 — 24 Categories

- Automobile and other motor vehicle dealers
- Automobile dealers
- Automotive parts, accessories, and tire stores
- Building material and garden equipment and supplies dealers
- Building material and supplies dealers
- Drinking places
- Floor covering stores
- Food and beverage stores
- Food services and drinking places
- Full service restaurants
- Furniture stores
- Gasoline stations
- Grocery stores
- Hardware stores
- Health and personal care stores
- Limited service eating places
- Motor vehicle and parts dealers
- New car dealers
- Office supplies and stationery stores
- Paint and wallpaper stores

- Pharmacies and drug stores
- Restaurants and other eating places
- Supermarkets and other grocery (except convenience) stores
- Used car dealers

Cluster 3 — 4 Categories

- Department stores (excluding discount department stores)
- Gift, novelty, and souvenir stores
- Hobby, toy, and game stores
- Jewelry stores

References

Bandara, K., Bergmeir, C., and Hewamalage, H. (2020). Forecasting hierarchical and grouped time series with high-dimensional predictors: A machine learning approach. *International Journal of Forecasting*, 36(4):1342–1356.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, pages 359–370.

Blinder, A. S. (2001). The macroeconomics of consumer behavior. *Brookings Papers on Economic Activity*, 2001(1):1–63.

Bontempi, G., Ben Taieb, S., and Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. *European Business Intelligence Summer School*, pages 62–77.

Carroll, C. D. (1997). Buffer-stock saving and the life cycle/permanent income hypothesis. *Quarterly Journal of Economics*, 112(1):1–55.

Carroll, C. D. and Summers, L. H. (1994). Consumption, income, and interest rates: Reinterpreting the time series evidence. *Carnegie-Rochester Conference Series on Public Policy*, 40:59–90.

Chen, Z., Liu, S., Chang, Z., Wu, Y., and Shi, K. (2022). Developing improved time-series dmsp-ols-like data (1992–2019) in china by integrating dmsp-ols and snpp-viirs. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19.

Dekimpe, M. G. and van Heerde, H. J. (2023a). Managing sales elasticities in times of inflation: A strategic marketing perspective. *International Journal of Research in Marketing*, 40(1):88–108.

Dekimpe, M. G. and van Heerde, H. J. (2023b). Retailing in times of soaring inflation: What we know, what we don't know, and a research agenda. *Journal of Retailing*, 99(3):322–336.

Federal Reserve Bank of St. Louis (2024). Fred economic data. `https://fred.stlouisfed.org`. Accessed: 2025-07-30.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hasan, M. et al. (2022). Retail sales forecasting: A comparison of arima, prophet, and lightgbm on walmart data. *arXiv preprint arXiv:2205.01432*.

Hobor, P. et al. (2025). Tree-based ensembles versus deep learning for retail sales prediction with macroeconomic features. *arXiv preprint arXiv:2502.01455*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, 3 edition. Accessed online at `https://otexts.com/fpp3/`.

Inayati, S., Iriawan, N., and Irhamah (2024). A markov switching autoregressive model with time-varying parameters. *Forecasting*, 6(3):568–590.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Ma'ady, M. N. P., Meilanitasari, P., Vanany, I., and Isrofi, N. (2023). Optimizing clustering techniques for retail industry: One-dimensional time warping method for demand uncertainty. In *International Conference on Sustainable Information Engineering and Technology (SIET)*, Badung, Bali, Indonesia. ACM.

Murry, J. and Vu, K. (2020a). Inflation and promotional timing: How consumers adjust purchase and stockpiling behavior. *Journal of Marketing Research*, 57(6):1054–1073.

Murry, J. and Vu, K. (2020b). Inflation and promotional timing: How consumers adjust purchase and stockpiling behavior. *Journal of Marketing Research*, 57(6):1054–1073.

Petitjean, F., Ketterlin, A., and Gançarski, P. (2011a). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Petitjean, F., Ketterlin, A., and Gançarski, P. (2011b). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Ranaweera, C. and Prabhu, J. (2021a). How inflation impacts private-label brand share: Evidence from consumer packaged goods. *Journal of Retailing*, 97(4):463–477.

Ranaweera, C. and Prabhu, J. (2021b). How inflation impacts private-label brand share: Evidence from consumer packaged goods. *Journal of Retailing*, 97(4):463–477.

Rasul, K., Schäfer, J., Bergmann, U., Dey, A., Zhang, X., Binas, J., Dürr, O., and Salinas, D. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv preprint arXiv:1912.09363*.

Shamim, A. (2023). Economic indicators and inflation. `https://www.kaggle.com/datasets/adilshamim8/economic-indicators-and-inflation`. Accessed: 2025-07-30.

Silva, J. E. S., Conceição, J. S. d., Costa, L. C. C. d., Fiaccone, R. L., Barreto, M. E., Ichihara, M. Y., and Ara, A. (2024). K-means dtw barycenter averaging: a clustering analysis of covid-19 cases and deaths on the brazilian federal units. *International Journal of Data Science and Analytics*.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.