The Hidden Complexity in Retail Demand

Michael Ha, Jee Hun Hwang, and Lei Lin

Shiley-Marcos School of Engineering, University of San Diego

ADS 506: Applied Time Series Analysis

Professor Dave Hurst

December 8, 2025

Github

Abstract

Retail demand forecasting is often simplified through aggregated modeling approaches, despite meaningful differences in demand behavior across products. This project examines four years (2014–2017) of retail transaction data across 17 sub-categories within the Office Supplies, Furniture, and Technology product categories to evaluate the limitations of uniform forecasting strategies. Exploratory analysis reveals substantial heterogeneity in trend, seasonality, volatility, and intermittency across sub-categories, even within the same parent category. To address this variability, sub-category–level forecasting models were developed using time series linear models (TSLM), exponential smoothing methods (ETS), and autoregressive integrated moving average (ARIMA) frameworks, with variance-stabilizing transformations applied as needed.

Models were trained on historical data and evaluated using a one-year holdout period with RMSE, MAE, and MAPE as performance metrics. Results demonstrate wide variation in forecastability: highly seasonal sub-categories such as Phones achieved MAPE values below 20%, while intermittent, capital-driven demand in Copiers exceeded 40% regardless of model complexity. Automated model selection supported consistent evaluation, while tailored specifications improved accuracy relative to aggregated approaches. Overall, the findings show that disaggregated, data-driven forecasting materially improves interpretability and forecast performance, highlighting the operational value of aligning inventory planning strategies with the underlying demand characteristics of individual product sub-categories.

The Hidden Complexity in Retail Demand

Retail businesses encounter an enduring challenge in inventory management, as various products demonstrate fundamentally divergent demand behaviors. Nevertheless, organizations frequently employ uniform forecasting methodologies across their entire product assortment. This disparity between product heterogeneity and standardized management practices results in tangible financial repercussions. Overestimations of demand lead to excess inventory, thereby immobilizing working capital and accruing warehousing and holding costs. Conversely, underestimations cause stockouts, culminating in lost sales and customer dissatisfaction, potentially prompting consumers to seek alternatives among competitors. The magnitude of these costs is heavily contingent upon forecast accuracy, the capacity to predict future demand and accordingly adjust inventory levels

Despite the high stakes, many organizations default to a deceptively simple approach by aggregating products into broad categories and applying standardized forecasting models uniformly across all items within each category. However, research on inventory coordination indicates that uniform strategies often increase overall costs, especially when products have diverse characteristics (Wang et al., 2011). This one-size-fits-all methodology offers undeniable administrative advantages. Fewer models require less maintenance. Planning processes remain straightforward. Reporting stays standardized. For organizations managing hundreds or thousands of products, the appeal of simplicity is compelling.

However, this simplicity comes at a steep cost when the fundamental assumption underlying aggregated forecasting proves false. The assumption that products with the same category label act similarly is rarely true in practice. Within any category, products often show

unique demand patterns caused by different underlying business processes. Some products exhibit pronounced seasonal patterns tied to retail calendars or weather patterns. Others follow corporate budget cycles and procurement schedules. Still others trend steadily upward with market growth, while others spike irregularly in response to external events, promotions, or competitive dynamics.

This project examines four years (2014-2017) of retail transaction data across 17 product subcategories within three major categories to explore demand differences within each category. The analysis reveals that products grouped within single categories exhibit dramatically different demand behaviors. By disaggregating demand patterns to the product level and developing tailored forecasting models for each product's unique characteristics, it is demonstrated that forecast accuracy and inventory efficiency can be significantly improved compared to conventional aggregated approaches.

Exploratory Data Analysis

The exploratory data analysis (EDA) was conducted in RStudio, using R coding language with tidyverse and fpp3 workflows to organize, summarize, and visualize the data. The dataset consists of 9,994 transactions spanning a continuous 48-month period from January 2014 to December 2017, representing 37,873 total units sold across three major product categories, Office Supplies, Furniture, and Technology. Within these categories, 17 distinct sub-categories contribute varying levels of demand, with Office Supplies accounting for the largest share of unit volume as in Figure 1, followed by Furniture and Technology. In Figure 2, at the sub-category level, groups such as Binders, Paper, Furnishings, and Phones consistently exhibit higher activity, highlighting early differences in product movement across the portfolio. Because the

objective is to understand underlying demand rather than revenue fluctuations, the analysts modeled Quantity sold rather than Sales, as Sales can be distorted by pricing changes and promotions, whereas Quantity provides a cleaner and more stable signal for inventory planning and forecasting.

Figure 1

Total Quantity Sold by Category

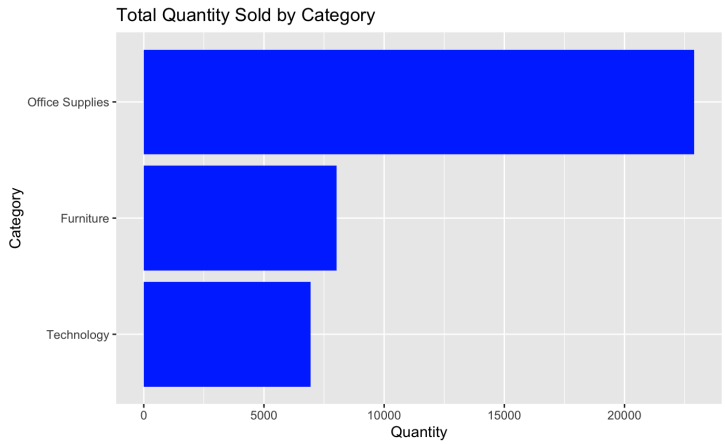
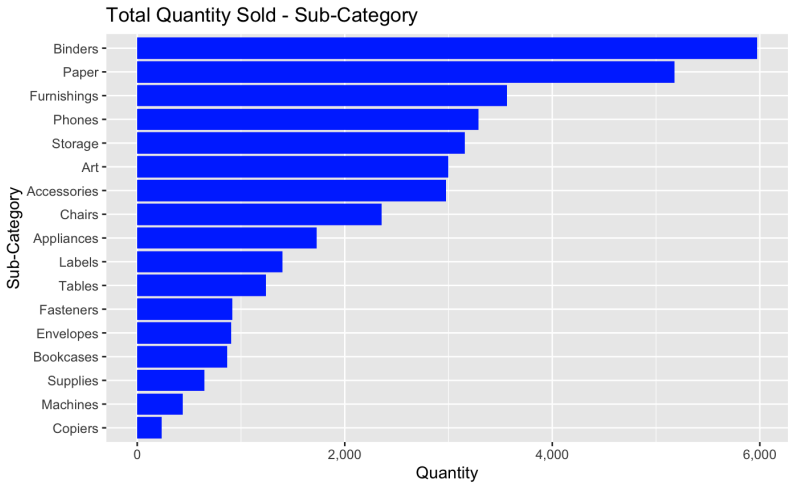


Figure 2

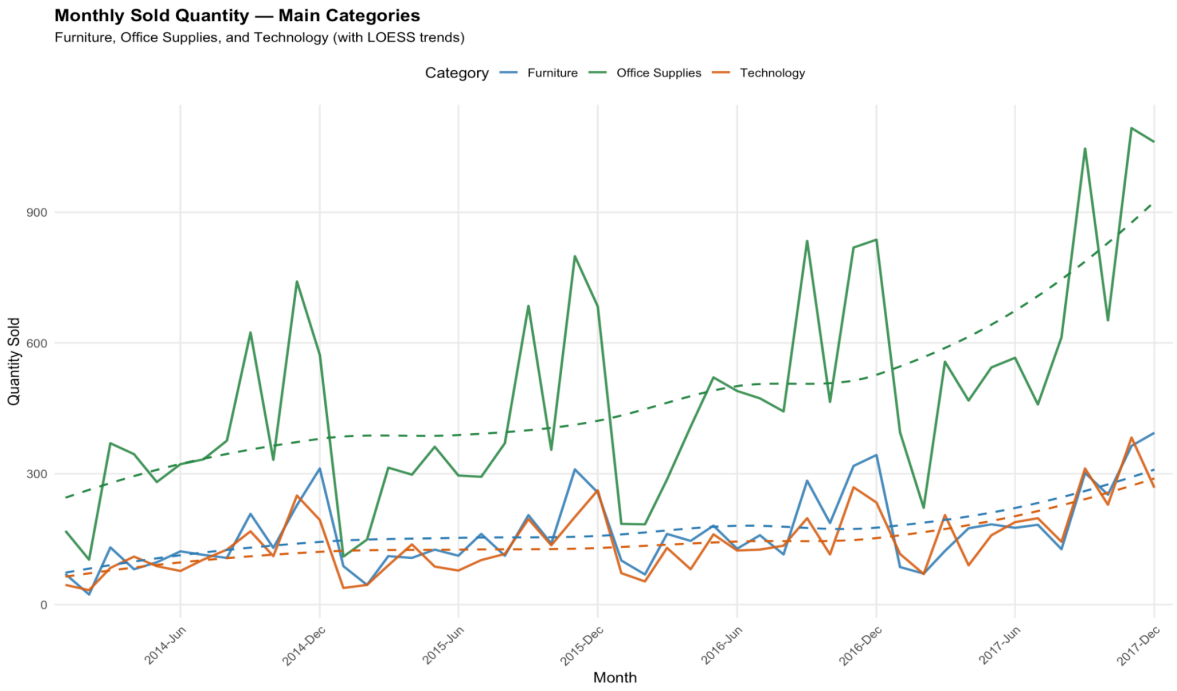
Total Quantity Sold by Sub-Category



When visualizing the time series of monthly quantities sold across the three main categories, Furniture, Office Supplies, and Technology, in Figure 3, distinct demand patterns begin to emerge. Office Supplies show both the highest volume and the greatest volatility, with recurring spikes and a sharply rising trend that may reflect routine business replenishment cycles and purchasing tied to organizational workflows. In contrast, Furniture and Technology follow more moderate and steady upward trajectories, with Technology growing at a rate similar to Furniture despite exhibiting slightly noisier short-term fluctuations, which may be influenced by product launches or tech refresh cycles. These steadier patterns likely stem from less frequent, higher-consideration purchases compared to Office Supplies' recurrent buying behavior. The differing shapes of these trend lines highlight that each category is influenced by unique underlying demand drivers.

Figure 3

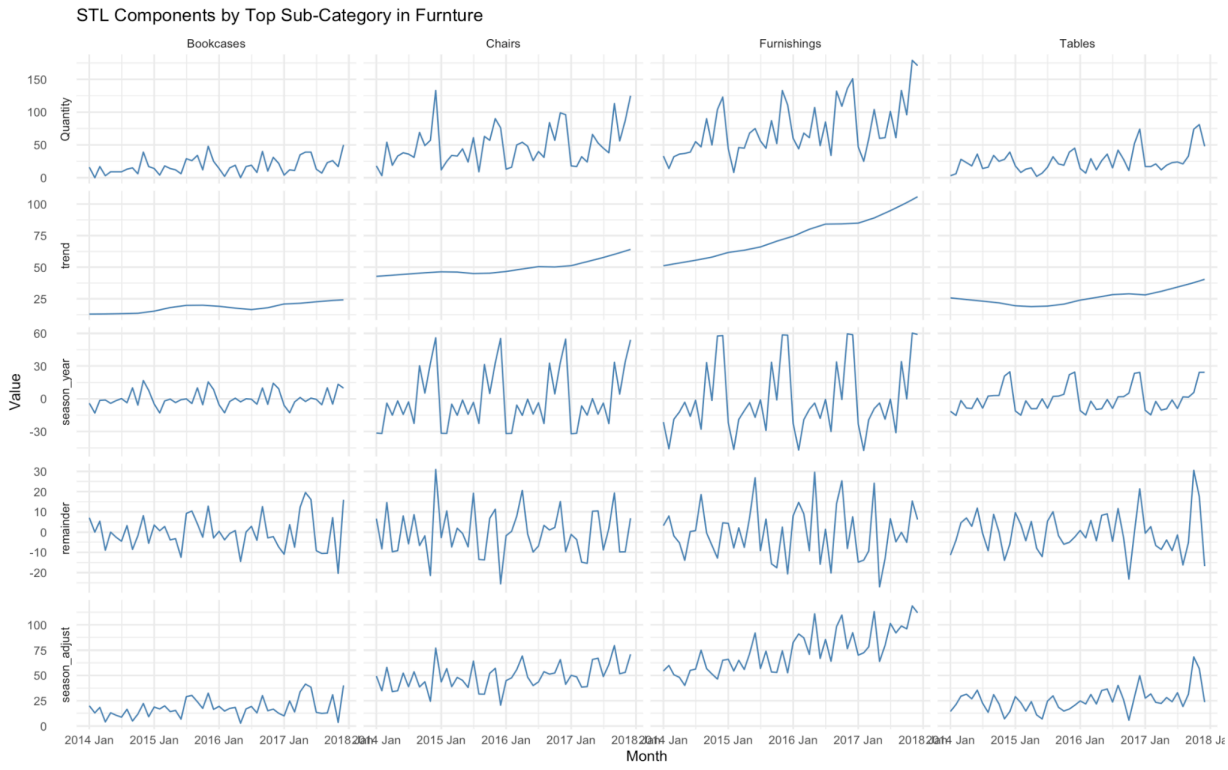
Time Series of Quantity Sold Monthly of the Three Main Categories



To further illustrate how demand patterns diverge within a single category, the STL decomposition was conducted on the four major Furniture sub-categories, Bookcases, Chairs, Furnishings, and Tables, as shown in Figure 4. The results show distinctly different structures in their trend, seasonal, and irregular components. Bookcases exhibit low and steady demand with minimal seasonality, while Chairs display clearer seasonal peaks and a gradually rising trend. Furnishings stand out with the strongest growth and the most pronounced recurring seasonal pattern, indicating a sub-category that is both expanding and highly cyclical. In contrast, Tables maintain relatively stable demand with only mild trend changes and modest seasonality. These contrasting behaviors demonstrate that even closely related products share little resemblance in their underlying time-series properties.

Figure 4

STL Components by Top Sub-Category in Furniture



Data Cleaning and Preparation

Prior to modeling, several data cleaning and preparation steps were conducted to ensure the dataset was suitable for time-series analysis. The transaction-level data was first aggregated into a monthly tsibble, summarizing total quantities sold for each Category and Sub-Category to establish a consistent temporal structure. Dates were reformatted into a yearmonth index, and an assessment of the dataset confirmed that no missing time periods or quantity values were present. During preparation, each sub-category was evaluated for variance instability, trend behavior, and seasonal structure, applying appropriate transformation techniques, including Box–Cox and other variance-stabilizing methods, when diagnostics indicated the need for adjustment. Seasonal or regular differencing was introduced only when required to address non-stationarity in the series. Finally, the dataset was partitioned into a training period (2014–2016) and a validation period (2017) to support reliable out-of-sample performance evaluation. Through these steps, each sub-category series was standardized into a clean and stable format suitable for modeling and forecasting.

Modeling

Category — Furniture

In modeling the Furniture sub-categories, complementary approaches such as TSLM, ETS, and ARIMA were adopted to address the distinct trend and seasonal structures observed in the data. The TSLM model relied on time-series–derived predictors, seasonal indicators and trend components, because these predictors can be projected into future periods without requiring external information, which is typically unavailable in ex-ante forecasting (Hyndman & Athanasopoulos, 2021). Transformations such as Box–Cox and other variance-stabilizing

techniques were applied when appropriate to reduce heteroskedasticity and better capture nonlinear patterns while maintaining linear model estimation. The ETS framework was used to model level, trend, and seasonality through exponentially decaying weights on past observations, enabling the model to adapt more heavily to recent demand shifts. ARIMA provided an alternative structure by leveraging autocorrelation and lagged relationships, reflecting the realistic notion that changes in demand often influence subsequent periods with some delay. These methods offered complementary strengths, which were also evaluated alongside ensemble strategies to incorporate the collective predictive information from TSLM, ETS, and ARIMA when beneficial.

For each Furniture sub-category, an automated modeling workflow was applied using the auto implementations of TSLM, ETS, and ARIMA to identify the best-fitting specifications based on information criteria and residual diagnostics. These automated procedures were selected because they provide a systematic, unbiased, and fast way to search across many possible model structures, reducing the risk of manual selection bias in the initial modeling and ensuring consistent evaluation across sub-categories. During this process, various transformations, including log1p to accommodate zero values, square-root transformations, and Box–Cox adjustments, were explored to stabilize variance and improve model performance when appropriate. Each transformed series was then evaluated across the three modeling frameworks, and the resulting models were compared using accuracy metrics on the validation periods. In addition to the individual models, ensemble variants combining forecasts from TSLM, ETS, and ARIMA were also assessed to examine whether pooled predictions offered improvements in stability or accuracy. This systematic procedure ensured that each sub-category

was matched with the most suitable model form based on empirical performance rather than assumptions.

Category — Office Supplies

For the nine Office Supplies sub-categories, ARIMA was selected as the sole modeling method due to its interpretability and strong ability to model monthly retail demand with seasonal patterns (Fattah et al., 2018). Both auto ARIMA and manually-tuned ARIMA models were developed, allowing automated parameter search to establish baselines while ACF/PACF diagnostics guided refinement when seasonal or autocorrelation patterns suggested alternative specifications. Seasonality was strongly present across these sub-categories with recurring annual peaks, resulting in the application of differencing strategies including non-seasonal $d = 1$, seasonal $D = 1$ (period 12), and combined $D12d1$ forms to achieve stationarity. Most of these sub-categories were stabilized using log transformation, but a few select required Box-Cox transformation after log-based models failed to meet stationarity thresholds during testing.

Residual performance served as the primary criterion for model selection between untransformed and transformed sub-category data, which was validated using Ljung-Box tests with sufficient lags to confirm independence and white-noise characteristics. In many cases, such as the Art and Binders sub-categories, transformation improved diagnostics; in others, such as the Appliances sub-category, raw models performed adequately. This highlighted the necessity of evaluating each sub-category individually rather than applying a blanket transformation strategy. Final models were chosen based on the combination of transformation, differencing, and ARIMA structure that produced stable forecasts based on accuracy metrics. This approach

resulted in consistent methodology across the Office Supplies sub-categories while still tailoring models to their unique seasonal dynamics.

Category — Technology

To model the Technology sub-categories, TSLM, ETS, and ARIMA were applied to address the substantial heterogeneity in demand behavior across Phones, Accessories, Machines, and Copiers. Demand patterns ranged from highly seasonal and predictable in Phones, driven by product launch cycles and holiday demand, to highly irregular, intermittent in Copiers, typical of capital-intensive purchases. TSLM captured trend and seasonal structure using time-derived predictors suitable for ex-ante forecasting (Hyndman & Athanasopoulos, 2021), while ETS offered adaptive smoothing that responded to recent shifts in consumer demand. ARIMA complemented these approaches by modeling temporal dependencies and lag effects. Variance-stabilizing transformations, which include log1p, square root, and Box-Cox, were applied when needed to reduce heteroskedasticity, especially in subcategories with uneven demand.

For each sub-category, automated TSLM, ETS, and ARIMA workflows were used to determine optimal model specifications based on information criteria and residual diagnostics. Models were assessed over a one-year holdout period using RMSE, MAE, and MAPE, revealing notable differences in forecastability: Phones achieved MAPE below 20%, while Copiers exceeded 40% across all models due to demand intermittency. Transformation strategies were evaluated alongside untransformed models, and ensemble forecasts combining all three methods were also evaluated, though the strongest individual models generally outperformed pooled predictions. Final model selection prioritized out-of-sample RMSE and diagnostic validity,

ensuring each Technology sub-category was paired with an empirically optimal forecasting approach rather than a uniform modeling assumption.

Results

Model Evaluation

Each model was trained using all historical data except the final year, then used to generate 12-month forecasts. Evaluation began by assessing whether differencing or transformation was required to achieve stationarity, and in many cases, the automatically tuned models (ARIMA, ETS, and TSLM) handled this without manual intervention. Transformations were also tested per series, including logarithmic (log, log1p), square-root (sqrt), and Box-Cox. Model accuracy was primarily assessed using RMSE, supplemented with ME, MAE, MPE, MAPE, MASE, RMSSE, and residual autocorrelation (ACF1) to ensure stability. Table 1 summarizes the best performing models per sub-category, including their forecast accuracy metrics and applied transformation type.

Table 1

Best Performing Models per Sub-Category

Sub-Category	Model	Type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1	Transform
Appliances	auto_v	Test	17.4	25.76	22.71	24.97	44.85	1.43	1.26	0.11	none
Art	auto_logv	Test	33.06	37.02	33.06	36.78	36.78	1.95	1.63	-0.15	log
Binders	arima_logv	Test	-37.65	70.2	55.94	-24.13	36.91	1.78	1.75	-0.05	log
Envelopes	arima_bcv	Test	-2.56	8.01	6.86	-45.8	62.38	0.68	0.71	0.05	boxcox
Fasteners	arima_bcv	Test	-38.91	46.75	38.91	-243.44	243.44	3.21	3.37	0.14	boxcox
Labels	arima_logv	Test	-2.88	24.13	18.27	-38.87	74.99	1.25	1.25	-0.05	log
Paper	arima_logv	Test	-7.42	41.35	35.11	-10.15	26.83	1.15	1.11	0.26	log
Storage	arima_logv	Test	-28.04	46.64	41.01	-43.61	58.55	1.6	1.54	-0.39	log
Supplies	arima_bcv	Test	-21.04	25.66	22.99	-197.69	209.13	2.54	2.22	0.11	boxcox

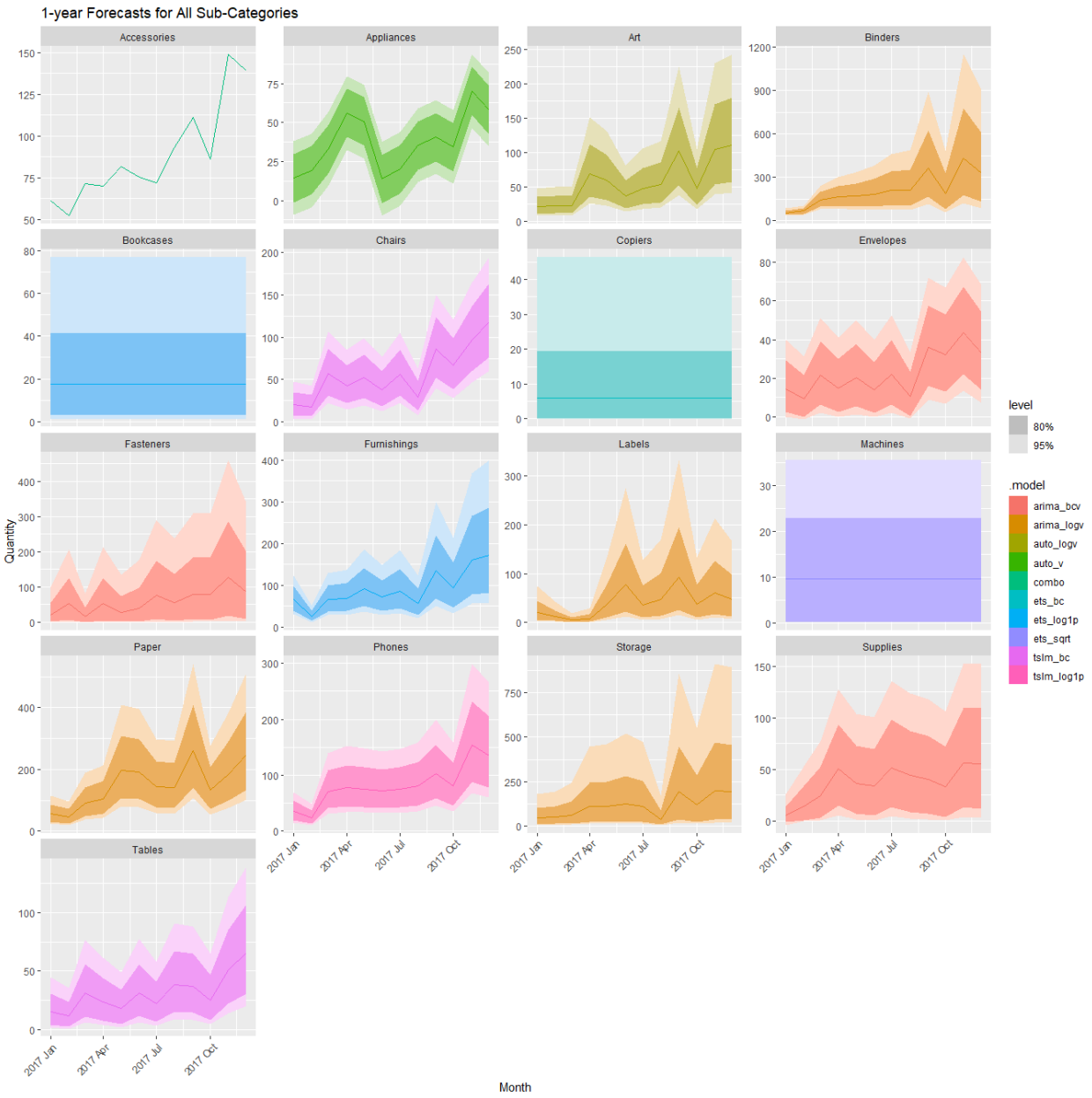
Bookcases	ets_log1p	Test	5.43	15.18	12.33	-29.51	76.28	1.57	1.54	0.19	log1p
Chairs	tslm_bc	Test	-0.71	14.65	12.52	-11.04	27.8	0.78	0.73	0.06	boxcox
Furnishings	ets_log1p	Test	0.52	16.58	11.97	-4.48	15.84	0.58	0.64	-0.34	log1p
Tables	tslm_bc	Test	1.51	18.82	13.2	-13.47	39.47	1.12	1.28	0.18	boxcox
Accessories	combo	Test	1.33	26.04	18.6	-21.4	35.25	1.09	1.22	-0.33	log1p
Copiers	ets_bc	Test	0.09	5.94	4.75	-Inf	Inf	0.94	0.9	-0.44	boxcox
Machines	ets_sqrt	Test	0.53	8.51	6.6	-126.38	158.12	0.68	0.68	0.35	sqrt
Phones	tslm_log1p	Test	8.65	24.61	22.08	2.64	32.61	1.2	0.99	-0.02	log1p

Forecasting Results

Figure 5 presents the 1-year ahead forecasts for all sub-categories, overlaid with 80% and 95% confidence intervals to explain variability. Each facet displays the forecasted trajectory for an individual sub-category, allowing visual comparison of both magnitude and seasonality between each product group. Overall, most forecasts show upward demand trends across the year with wider confidence bands in categories prone to volatility (e.g. Binders, Paper, Storage), and tighter prediction intervals for more stable series such as Appliances, Chairs, and Furnishings. The diversity in slope, amplitude, and uncertainty bands highlights why sub-category-specific modeling is necessary, rather than aggregated forecasting by category. The figure also confirms that models captured seasonal fluctuations and year-end increases where historically observed.

Figure 5

1-Year Forecasts for All Sub-Categories



To illustrate practical forecasting insights, the Office Supplies sub-categories were examined using the monthly mean forecast values, as observed in Table 2. Binders and Paper show the most pronounced growth, climbing from 54 and 56 units in January to peaks of 434 and

247 by November/December, while Storage also trends upward through late Q3 and Q4, suggesting strong seasonal demand. Conversely, Appliances and Envelopes remain relatively stable and low-volume throughout the year, indicating lower forecasting risk and slower turnover, and mid-range categories such as Fasteners, Labels, and Supplies exhibit periodic spikes that may align with late-summer and holiday promotions. These results collectively demonstrate how sub-category-level forecasting supports more informed purchasing and replenishment scheduling, in which high-growth segments like Binders and Paper may require more agile inventory strategies while stable groups like Appliances and Envelopes enable longer reorder cycles.

Table 2

Monthly Mean Forecasts for Office Supplies (12-Month Horizon)

Month	Appliances	Art	Binders	Envelopes	Fasteners	Labels	Paper	Storage	Supplies
2017 Jan	14	22	54	14	22	21	56	46	5
2017 Feb	19	23	66	9	54	12	46	50	14
2017 Mar	33	23	143	22	15	5	91	63	24
2017 Apr	56	69	167	15	54	8	103	110	51
2017 May	51	60	171	20	29	37	197	110	36
2017 Jun	14	37	184	14	40	77	190	123	34
2017 Jul	20	48	208	22	76	36	144	110	51
2017 Aug	35	53	211	10	57	48	141	39	44
2017 Sep	41	103	365	36	79	93	261	193	40
2017 Oct	34	48	188	32	78	37	133	122	33
2017 Nov	70	105	434	44	129	60	184	200	56
2017 Dec	58	111	330	33	85	47	247	192	55

Discussion

The findings from this project highlight both the advantages and the limitations of classical time-series forecasting methods when applied to heterogeneous retail demand. While sub-category–specific modeling substantially improved accuracy relative to aggregated approaches, the results also revealed opportunities for methodological refinement. For the TSLM framework, future work could incorporate regression splines or piecewise linear trend structures to capture mild nonlinearities observed in several sub-categories. These approaches would allow localized changes in trend, such as accelerations during growth phases or plateaus in mature product lines, without abandoning the interpretability of linear models. However, such flexibility also carries a heightened risk of overfitting and may introduce bias if knot locations are not selected cautiously (Hyndman & Athanasopoulos, 2021). Another promising direction involves identifying and integrating causal predictors. If confounding variables such as promotional intensity, macroeconomic indicators, or organizational procurement cycles can be quantified and reasonably forecasted, they could enhance ex-ante forecasting by providing signals beyond what is available in the time-series patterns alone.

Within the ETS framework, additional variants merit exploration, particularly damped trend models, which may better reflect realistic long-term behavior by preventing trends from extrapolating indefinitely. Similarly, choosing between additive and multiplicative seasonal components could improve generalization, as some sub-categories exhibit seasonal amplitudes that scale with their level (Hyndman & Athanasopoulos, 2021). Evaluating such variants systematically may reveal ETS specifications better aligned with the evolving seasonal structures observed in categories like Furnishings and Phones.

For ARIMA models, the results showed strong capability in modeling short-term autocorrelation and annual seasonality; however, future iterations could incorporate alternative seasonal differencing strategies or hybrid ARIMA-ETS architectures to address sub-categories with both strong autocorrelation and volatile seasonal amplitude (Hyndman & Athanasopoulos, 2021).

Beyond methodological enhancements, the study underscores the importance of understanding the operational context behind each time series. Sub-categories with high intermittency or sporadic large orders, such as Copiers, may require models specifically designed for intermittent demand, such as Croston's method or probabilistic count-based forecasting (Hyndman & Athanasopoulos, 2021).

At a strategic level, integrating forecasting outputs into a demand calendar can provide cross-functional teams with clearer visibility into anticipated demand cycles, enabling improved planning for inventory, marketing campaigns, and budgeting. Finally, establishing an automated, rolling re-training pipeline would ensure that model parameters remain responsive to emerging demand shifts and external disruptions.

Overall, the analysis demonstrates that while classical forecasting models provide a strong foundation for sub-category–level forecasting, meaningful improvements can be achieved by expanding model flexibility, incorporating causal structure, exploring specialized methods for irregular demand, and embedding forecasts within operational decision-making workflows. In addition, modern approaches such as XGBoost and neural network–based models offer the ability to capture nonlinear relationships and complex interactions that traditional methods may overlook. Emerging large time-series models like TimeGPT also present new opportunities for

scalable, pretrained forecasting that can adapt across diverse retail sub-categories with minimal tuning. Further enhancements could be achieved through more rigorous cross-validation (CV) strategies, such as rolling-origin or sliding-window CV, to ensure that model performance remains stable and generalizable over varying demand conditions.

Conclusion

In conclusion, this project demonstrates that retail demand forecasting benefits substantially from disaggregated, sub-category–level modeling rather than aggregated approaches that assume homogeneous behavior within product categories. Exploratory analysis revealed pronounced differences in trend, seasonality, volatility, and intermittency across the 17 sub-categories, and the modeling results confirmed that no single forecasting method consistently outperformed others across all series. Instead, the most accurate forecasts emerged when model selection, transformations, and differencing decisions were tailored to the structural characteristics of each sub-category. Through the adoption of TSLM, ETS, ARIMA, and ensemble methods, individualized forecasting strategies were shown to improve predictive accuracy while also producing insights with direct operational value for inventory planning, promotional timing, and budgeting. These findings highlight the importance of aligning forecasting techniques with the underlying dynamics of specific product groups and lay the groundwork for future enhancements incorporating causal predictors, nonlinear trend modeling, and automated model governance systems.

References

Berry, L. R., Helman, P., & West, M. (2020). *Probabilistic forecasting of heterogeneous consumer transaction-sales time series. International Journal of Forecasting.* Advance online publication. https://doi.org/10.1016/j.ijforecast.2019.09.017

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). *Forecasting of demand using ARIMA model. International Journal of Engineering Business Management, 10,* 1–9. https://doi.org/10.1177/1847979018808673

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Pavlyshenko, B. M. (2019). *Machine-Learning models for sales time series forecasting. Data, 4*(1), 15. https://doi.org/10.3390/data4010015

Wang, K.-J., Lin, Y. S., & Yu, J. C. P. (2011). *Optimizing inventory policy for products with time-sensitive deteriorating rates in a multi-echelon supply chain. International Journal of Production Economics, 130*(1), 66–76. https://doi.org/10.1016/j.ijpe.2010.11.009