A statistical approach to analyzing and forecasting retail demand

by

Tinashe Sekabanja

B.S., Kansas State University, 2025

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Dr. Perla E. Reyes

Copyright

Abstract

Retail analytics enables businesses to analyze purchasing patterns, forecast demand, and improve inventory management. This study explores and compares statistical and machine learning methods for modeling retail demand using a synthesized dataset from Kaggle, the Product Sales Dataset (2023–2024) by Yennewar (2023). The dataset included transaction records across multiple states, product categories, and time periods in the United States.

Exploratory analysis was conducted to examine geographic and temporal patterns in product demand, revealing significant regional variations that were confirmed through analysis of variance (ANOVA). Transaction-level data were aggregated to monthly and quarterly product-level demand by state and city, and missing combinations were reconstructed to account for no purchases. Thus, producing a complete dataset suitable for analysis.

To model demand, a combination of time series, statistical, and machine learning approaches were implemented. ARIMA models were used to capture temporal trends, while count based regression models (Poisson and Negative Binomial) were applied to model demand. Additionally, classification methods including ordinal logistic regression, decision trees, and XGBoost were applied to predict categorized demand levels. Model performance was evaluated using metrics such as mean absolute error and accuracy.

Results indicated that demand varies significantly across both regions and time. Product-level time series models captured underlying trends less effectively at aggregated levels. In the regression approach, the Negative Binomial model outperformed the Poisson model due to overdispersion in the data. Classification models showed moderate performance with stronger predictive results for extreme demand categories (very low demand and high demand). Notably, XGBoost underperformed compared to the other models but had a uniform moderately weak performance across all classes.

Overall, the findings highlight the importance of aligning model selection with data structure and distribution as well as problem structure, demonstrating that increased model complexity does not necessarily lead to improved predictive performance in forecasting retail demand.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my major professor, Dr.Perla Reyes for her support and guidance throughout this study. Additionally, I would like to thank the other professors on my committee, Dr. Michael Higgins and Dr. Christopher Vahl. Dr. Reyes, Dr. Higgins, and Dr. Vahl have all been influential in shaping my statistical passions and journey from the undergraduate to masters level. I would also like to thank all the faculty in the department and in the university who have contributed to my academic and professional journey. Lastly, I would like to thank my friends and family for their continuous support.

Dedication

This report is dedicated to my parents, Mr. and Mrs. Sekabanja. They watered the seeds of my early love for numbers which have since blossomed into a passion for statistics.

Chapter 1

Introduction

Retail analytics is commonly defined as the use of software and analytical tools to collect and analyze data from physical, online, and catalog sales channels to generate insights into customer behavior and shopping trend (Hickins, 2023). These insights enable retailers to identify purchasing patterns, forecast demand, and make informed decisions related to pricing strategies, promotions, staffing, and facility designs. When applied effectively retail analytics can support revenue growth while maintaining operational efficiency and profitability.Retail demand forecasting focuses on leveraging historical data, market trends, and advanced analytics to project consumer demand for a product. This allows retailers to maintain optimal inventory levels and predict trends in sales. Demand forecasting is integral as it helps businesses minimize overproduction, excessive inventory and emergency procurement costs, resulting in improved supply chain efficiency (Tredence, 2024).

This study analyzes a synthesized Product Sales Dataset (2023–2024) retail transactions dataset containing transaction records across multiple states, product categories and time periods in the United States from Kaggle (Yennewar, 2023). We were limited by the inability to obtain real data as they tend to be either proprietary or very expensive. The original transactional level data was aggregated to a product monthly level grouped by state and city. To ensure completeness of the data, missing combinations of geographic and temporal variables were reconstructed to account for geographic and temporal points where a product

was not purchased. This resulted in a dataset that was appropriate for subsequent analysis.

Exploratory data analysis was conducted to examine geographic patterns in demand which revealed significant regional variation that was further confirmed by analysis of variance (ANOVA). Based on these findings, all following analysis was performed on the West region. Principal Component Analysis (PCA) was used to explore seasonality in demand across months.

This study aims to support product demand planning for a synthetic retail company, specifically to optimize warehouse stocking decisions across states over the period of a year. A quarterly modeling approach was adopted to assist business leaders in managing inventory for non-perishable items. A key focus of this study was to predict the demand for specific products across different time periods.

Time series models were used to capture trends in individual products and forecast future demand. It utilizes generalized linear models, including Poisson and Negative Binomial regression to model demand as a function of explanatory variables. Ordinal logistic regression was also applied to predict demand categories. Additional predictive methods like classification trees, and extreme gradient boosting (XGBoost) were implemented to compare traditional statistical approaches with modern machine learning approaches.

Overall, this study demonstrates how a combination of statistical modeling, machine learning methods, and time series analysis can provide a comprehensive framework to analyze, forecast, and classify retail demand which highlights the value of data driven analytics in the retail industry.

Chapter 2

Methods

2.1 Introduction

This section outlines the modeling approaches used to analyze retail demand in this study. There were three main approaches used: time series forecasting, count based regression/ linear models, and classification modeling. Each approach was selected to model different aspects of demand. Time series models were used to capture temporal patterns of each product and generate forecasts. Count based regression models were used to model demand as a function of explanatory variables, accounting for the discrete nature of quantity demanded. Classification models were also implemented to categorize demand levels for inventory planning. These methods provide a comprehensive framework for forecasting, predicting, and classifying demand in the dataset.

2.2 Time Series Modeling

A time series is a sequence or a continuous collection of observations taken sequentially over time. In this study, demand data was modeled by both monthly and quarterly frequencies to evaluate two possible forecasting scenarios.

Statistical inference for time series is usually based on a single realization of the stochastic

process. That is, to use the observed values $y_1, y_2, \ldots, y_T$ over $t = 1, 2, \ldots, T$ to estimate/ infer some of the probability characteristics of the stochastic process $(Y_1, Y_2, \ldots, Y_T)$ that can be used to analyze the process and forecast (predict) future time points.

It is impossible to obtain an "accurate" estimate of the complete joint probability distribution of $Y_1, Y_2, \ldots, Y_T$, unless we make very strong assumptions.

Time series models can be reduced to models, with time t as the only predictor variable. Key aspects of time series models include trend, seasonality and the stationarity assumption. Trend refers to a long term increase or decrease in data while seasonality refers to a pattern that occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. It is important to note that seasonality is always of a fixed and known frequency. The stationarity assumption refers to a time series whose statistical properties do not depend on the time at which the series is observed, assuming constant mean and variance over time. Differencing, however, is a technique that can be used to make non-stationary time series data stationary by eliminating or reducing trends and seasonality (Hyndman and Athanasopoulos, 2018).

There are four main types of models: the Auto Regressive (AR Model), the Moving Average (MA model), the Autoregressive and Moving Average (ARMA model), and Autoregressive Integrated Moving Average (ARIMA model). ARIMA models were used in this study. ARIMA models are a combination of autoregression and a moving average model that create a non-seasonal ARIMA model. The full model is written out as:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Where y'_t is the differenced series that could have been differenced multiple times. The predictors are composed of y_t and lagged errors. An ARIMA(p, d, q) model consists of parameters $p =$ order of the autoregressive part, $d =$ degree of the initial differencing involved, and $q =$ order of the moving average part. It is important to note the special cases of ARIMA models: White noise: ARIMA$(0, 0, 0)$, Random walk ARIMA$(0, 1, 0)$ with no constant, Random walk with drift ARIMA$(0, 1, 0)$ with a constant, Autoregression ARIMA$(p, 0, 0)$, and

Moving average ARIMA$(0, 0, q)$.

Model selection was performed using the `auto.arima()` function in R. This function performs a stationarity check that determines d by applying differencing to reach stationarity. It then computes ACF and PACF to provide initial guesses for p and q, using a stepwise algorithm to test combinations of parameters. Lastly, the goodness of fit is evaluated using AIC, AICc, or BIC values, criteria that balance model complexity and goodness of fit. With the optimal parameters identified, the model is fit using maximum likelihood estimation (MLE).

2.3 Count-Based Regression/Linear Models Approach

The objective of the count-based regression approach, was to model product demand as a function of some explanatory variables while accounting for the discrete nature of demand. Since demand is measured by the quantity of units sold, it is represented by non-negative integer data, making count-based models appropriate for the analysis.

Poisson regression was selected as an initial baseline due to its use in modeling count data. For Poisson models, variance increases with the mean hence variance approximately equals the mean value. Under this framework, demand is then modeled as a function of predictor variables through a log link function.

However, further analysis of the data revealed overdispersion in the model. Overdispersion occurs when the variance of the response is greater than its mean and is a violation of the equidispersion of the Poisson model. To address this limitation, a Negative Binomial regression model was utilized. The Negative Binomial model accounts for dispersion by introducing an additional parameter allowing the variance to exceed the mean. The added flexibility makes it more suitable for modeling overdispersed count data (Serra, nd).

The `glm.nb()` function in R is from the MASS package and is used to fit Negative Binomial Generalized Linear Models. `glm.nb()` fits a model to predict a count response variable, e.g., quantity demanded based on a collection of predictors, here, product name, category, location, and temporal factors were used as explanatory variables. Both models

were evaluated using an out of sample evaluation test.

2.4 Classification Modeling Approaches

The primary objective of this classification modeling approach was to predict quantity demanded for various products. Instead of predicting exact discrete demand values with high variability and sensitivity to noise, a classification based approach was adopted to group demand into meaningful categories. To achieve this, demand categories were created based on values derived from quantiles to ensure balanced representation across classes. This transformation of the data enables the subsequent models to focus on a classification prediction rather than precise highly variable numerical predictions. Three modeling approaches were implemented to predict demand categories based on product characteristics, as well as geographical and temporal features.

2.4.1 Ordinal logistic regression model

First, an ordinal logistic regression model was implemented using `polr()` function in R to account for the ordered nature of the response variable. Ordinal logistic regression creates a model that can be used to make cumulative probability predictions or to quantify how predictors are associated with cumulative probabilities (Pitawela et al., 2026). This is most appropriate when the response variable consists of ordered categories. A proportional odds model is defined as:

$$logitP(Y \leq k|\mathbf{X}) = \zeta_k - \eta$$

for $\zeta_0 = -\infty < \zeta_1 < \cdots < \zeta_{K-1} < \zeta_K = \infty$ with $\eta = \beta'\mathbf{X}$ the usual linear predictor and ζ_k ordered threshold parameters (Venables and Ripley, 2002). This formulation assumes that the effect of the predictors is constant across all cumulative splits of the response variable, an assumption known as the proportional odds assumption. This model was evaluated using an out of sample evaluation test.

2.4.2 Decision trees

Decision trees were used to capture nonlinear relationships between predictors without strong distributional assumptions. Tree based models are a class of nonparametric algorithms that operate by partitioning the feature space into a number of smaller (non-overlapping) regions with similar response values using a set of splitting rules. This study utilizes the classification and regression tree (CART) algorithm proposed in Breiman et al. (1984). Decision trees partition data into homogeneous subgroups then fit simple constants in each subgroup. These subgroups are formed recursively using binary partitions resulting in an inverted tree like structure. During the partitioning, the objective at each node is to find the best feature and partition point. This is done by minimizing node impurity, measured by the Gini index.The algorithm selects splits that result in the most homogeneous nodes. The Gini index to minimize is defined below:

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k represents the proportion of observations belonging to class k in the node. The growth of the tree is restricted by stopping criteria and pruning techniques to prevent overfitting. The model's performance was evaluated by testing its performance using an out of sample test.

2.4.3 XGBoost

Extreme Gradient Boosting (XGBoost) is a high-performance ensemble learning method. XGBoost builds a sequence of decision trees within a gradient boosting framework, where each successive tree is fit to the residual errors (gradients) of the previous model, iteratively improving predictive accuracy.

The model optimizes a regularized objective function consisting of a loss function and a complexity penalty:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.1}$$

where $l(\cdot)$ represents the loss function (multi-class log-loss in this case), and $\Omega(f_k)$ pe-

nalizes the complexity of the model (i.e., the regression tree functions). The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum w_j^2 \tag{2.2}$$

where T is the number of leaves and w_j are the leaf weights. This regularization helps prevent overfitting by punishing overly complex models (Chen and Guestrin, 2016).

The `multi:softprob` objective in XGBoost is used for this multi-class classification problem, as the target variable contains more than two classes. This objective function outputs probability estimates for each demand category. The target variable was encoded as integer labels starting from zero, as required by the XGBoost implementation (XGBoosting, 2023).

XGBoost is particularly well-suited for this study due to its ability to capture nonlinear relationships between predictors. Model performance was evaluated using a train-test split.

2.5 Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) was applied as a dimensionality reduction technique to simplify the structure of the dataset. In this study PCA was used to investigate seasonality in demand over the months.

PCA reduces data complexity by transforming the original set of potentially correlated variables into a smaller set of uncorrelated components. By projecting the data onto orthogonal principal components, the method retains most of the variance in the dataset while reducing redundancy among features. These components can be interpreted of as linear combinations of the original variables that are optimally weighted and derived from the correlation matrix of the data.

The first few principal components explain the largest proportion of the total variance and can be retained for further analysis. Additionally, PCA helps reduce the impact of noise by emphasizing directions of maximum variance (Jolliffe, 2002).

2.6 Model Evaluation Strategy

Model performance was assessed using a combination of metrics tailored to each modeling approach, allowing for a comprehensive comparison across forecasting, regression, and classification tasks.

For time series forecasting models, model performance was evaluated using standard forecasting accuracy metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Scaled Error (MASE).

The RMSE is calculated by taking the square root of the average squared prediction errors, placing a larger emphasis on larger errors, making it sensitive to outliers. A lower RMSE indicates a better performing model, however, the value has to be examined on the scale of the target variable for interpretation (Calzone, 2022).

In contrast, the MAE is calculated by taking the average of the absolute values of the errors. This gives both small and large errors equal weight, resulting in a less outlier sensitive metric (Calzone, 2022).

The MASE was used to provide a scale independent measure of forecast accuracy. The MASE scales prediction errors relative to a naïve forecasting benchmark allowing for forecast accuracy comparison across series with different units. The scaled error for non seasonal time series is defined as:

$$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|}$$

Where e_j represents the forecast error. MASE is then computed as:

$$\text{MASE} = \text{mean}(|q_j|)$$

MASE values less than one indicate a better forecast than the average one-step naïve forecast computed on the training data. Alternately, MASE values greater than one, indicate worse performance than the average one step naïve forecast computed on the training data (Hyndman and Athanasopoulos, 2021).

For count based regression models, the models were evaluated using Akaike Information

Criterion (AIC), MAE, RMSE, and Normalized MAE (NMAE). The AIC is a model selection tool calculated using the likelihood function and the number of parameters. AIC can be used to select between competing models with the lowest AIC being the best model and is calculated by.

$$\text{AIC} = 2p - 2\ln(L)$$

where, p is the number of predictor variables in the model, and L is the value of the likelihood function for the model in question.

For AIC to be optimal, n must be large compared to p (O'Brien, 2022).

MAE and RMSE were used to assess predictive accuracy, while Normalized MAE provides a scale independent measure of error by expressing the average absolute error relative to the mean of the observed demand (Amperon, 2023). The normalized MAE (NMAE) is defined as:

$$\text{NMAE}(y, \hat{y}) = \frac{\text{MAE}(y, \hat{y})}{\text{mean}(y)} \tag{2.3}$$

Where $\bar{y}$ represents the mean of the observed values.

For the classification models, performance was evaluated using overall accuracy and balanced accuracy. Accuracy is the proportion of all classifications that were correct, whether positive or negative (Google Developers, 2024). It is mathematically defined as:

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, represent true positives and true negatives and FP, FN represent false positives and false negatives.

Balanced accuracy computes the average recall across all classes. It is defined as:

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}$$

where K is the number of classes, and TP_k and FN_k represent the true positives and false negatives for class k, respectively Grandini et al. (2020).

These various methods of comparison enabled consistent comparison across models while accounting for the differing objectives of the forecasting, regression, and classification approaches.

Chapter 3

Analysis and Results

This dataset contains 200,000 observations, each representing an individual customer order. As shown in Table 3.1, it includes a combination of numeric variables (Order ID, Quantity, Unit Price, Revenue, and Profit), and categorical variables (Order Date, Customer Name, City, State, Region, Country, Category, Sub-Category, and Product Name), with geographic information spanning 4 Regions, 47 U.S. states and 108 cities. The dataset contains 4 major categories with 19 sub-categories showing a diverse mix of products. Customer identifiers contained 120,230 unique customer names, while on a product level, the dataset contained 49 unique products. Numerical variables showing purchasing behavior and financial performance. On average, customers purchased 1.85 units per order with a standard deviation of 1.10, indicating the data consists of a small number of items. The average unit price is $382.86 with a standard deviation of $276.87, indicating a wide range of product price points. Revenue for each transaction averages $712.04 while average profit is $157.74, both indicating high variability, suggesting heteroskedasticity in transaction prices and profitability.

Overall, the dataset combines multiple categorical features with continuous financial metrics.The combination of these variables makes it well suited for predictive modeling and analysis of retail demand patterns.

Table 3.1: *List of Variables from Kaggle's Product Sales Dataset (2023-2024) with Summary (Yennewar, 2023).*

Variable	Description	Summary
Order_ID	Unique identifier for each order	200,000 unique values
Order_Date	Date of transaction	Transaction dates spanning 2 years
Customer_Name	Name of the customer	120,230 unique values
City	City of the customer	108 unique values (e.g., Tampa, Omaha)
State	State of the customer	47 unique values (e.g., Oregon, Iowa, Texas)
Region	Region of the country	4 levels: East, West, South, Central.
Country	Country where purchase was made	United States
Category	Broad product category	4 levels (e.g., Accessories, Clothing & Apparel)
Sub_Category	Subdivision of category	19 levels (e.g., Sportswear, Storage, Bags)
Product_Name	Product description	49 unique products (e.g., Vase, Crocs, Belt)
Quantity	Units purchased	Mean = 1.85, SD = 1.10, Range = [1, 11]
Unit_Price	Price per unit (USD)	Mean = 382.86, SD = 276.87, Range = [17.03, 1432]
Revenue	Total sales (Quantity × Unit Price)	Mean = 712.04, SD = 742.47, Range = [17.03, 9014.25]
Profit	Net profit per transaction	Mean = 157.74, SD = 155.69, Range = [3.92, 2763.72]

3.1 Data Preprocessing

The original Kaggle dataset was imported and stored in a structured data frame. To simplify the temporal analysis, the date variable was decomposed into separate year and month components.

The analysis focus was on modeling and prediction of demand. Therefore, the individual transactions were aggregated at a granular level by Region, State, City, Year, Month, Category, Sub_Category, and Product_Name. This level of aggregation was chosen to preserve geographic, temporal, and product level variation. Identifier variables such as Order_ID and Customer_Name were no longer needed after aggregation as the modeling efforts for the

current report are aimed at the product level, not individual transactions.

Because the original dataset only contained records of purchases, there was no record of months where particular products had not been ordered. Following Gates (2025), to ensure data completeness, the dataset was reconstructed to include a complete representation of demand behavior . The new complete dataset contained all combinations of year, month, state, and city, including instances where a product was not sold. These observations that were previously unrecorded were entered as zero demand entries. This helped reduce the censoring caused by aggregating individual orders, by recovering the "zero" cases, a relevant category in demand prediction.

For the financial variables, we choose to drop Revenue and Profit as the primary focus of analysis was demand in terms of number of units and not profitability. Price information was included by aggregating mean unit price at the product level.

Additional versions of the dataset were constructed to carry on specific modeling tasks. To analyze temporal trends further, a second dataset was constructed by aggregating monthly observations into quarterly periods. This transformation reduces short term noise in the data. From a business perspective, quarterly aggregation is also meaningful as inventory planning for non-perishable or low frequency items often occurs on a less frequent basis to minimize shipping and logistical costs.

For classification based modeling, demand quantities were categorized into ordinal demand categories: High(29 or more), Medium(16-28), Low(9-15), and Very Low (0-8) based on quantile thresholds. This transformation aids demand based inventory planning and enables the application of classification models such as ordinal logistic regression and tree classification methods as they benefit from balanced classes (Sydow, 2020).

For model evaluation, the dataset was split into training (75%) and testing (25%) sets with the seed set at 123 for reproducibility.

Finally, to support dimensionality reduction and trend discovery, a third dataset was constructed where each product is represented by its demand across months. This structure enables the application of Principal Component Analysis (PCA) to identify key patterns, including months associated with relatively high or low demand for specific products.

These preprocessing steps ensured that datasets had the proper structure for the application of time series, regression, and classification models described in the sections below.

3.2 Exploratory Data Analysis

The exploratory data analysis revealed skewed distributions and regional patterns in demand for various products, along with other variables. Product demand distributions were positively skewed as shown in Figure 3.1, indicating the presence of high demand orders that would affect overall demand data.

Despite the similarity in distributions by categories (Fig. 3.1 (left)), a chart of purchased categories revealed unique trends by region with the East and West Regions having the highest demand (Fig. 3.3 and 3.2 (right)). This suggests that consumer preferences vary geographically. Bar charts of the demand by product category (Fig. 3.2 (left) and 3.3) showed electronics and home & furniture as the highest demand-generating category across regions, indicating that demand also varies categorically.

These differences reinforce the presence of region and category based demand structures, highlighting the importance of exploring regional and product level analysis to capture different trends in the following modeling. A key focus of this project was to understand how demand changes over time in different locations. We will explore regional and temporal trends in the following subsections.

3.2.1 Evaluating Regional Variability

To formally asses regional differences detected in Figure 3.1 (right), a one way ANOVA (classical F-test) was conducted to test the null hypothesis that mean demand is equal across all regions. However, the resulting p-values (all $p < 9.6 \times 10^{-15}$) provide strong evidence against the null hypothesis and indicate statistically significant difference. The ANOVA test revealed that the observed differences in demand between regions were far too large to be due to random chance confirming region plays an important role in modeling demand. Results

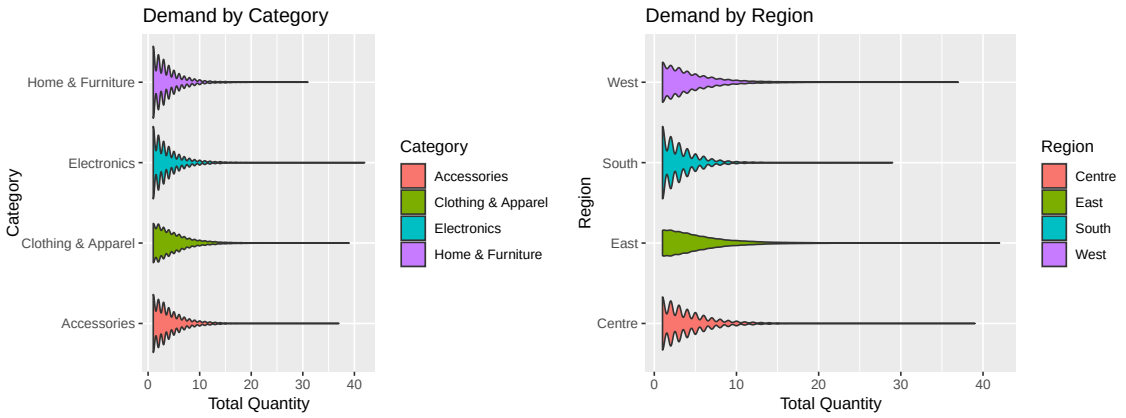

Figure 3.1: *Purchase Quantities per Transaction Distributions by Category (left) and Distributions by Region (right).*

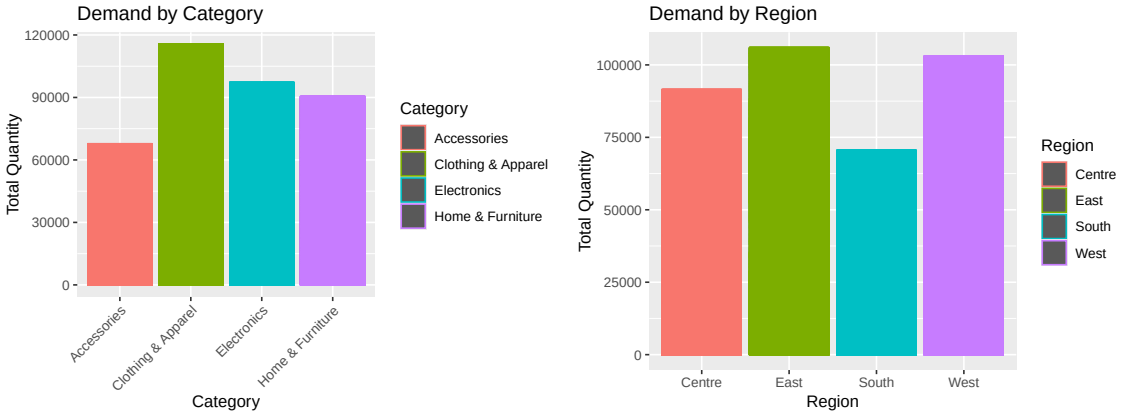

Figure 3.2: *Total Purchases by Category (left) and Region (right). The demand changes across regions and categories.*

are plotted in Figure 3.4

Results from the exploratory data analysis and ANOVA test indicated statistically significant differences in demand across regions. This suggests that region plays an important role in predicting demand. A general model for all regions was originally considered. However, the regional effect might go beyond an additive effect as illustrated by its interaction with product category (Fig. 3.3). Therefore, we opted to focus on determining an analysis strategy for the West region only and to develop a targeted simpler model that better reflects local demand trends.

The analysis strategy developed in the current report can be replicated for other regions. Due to time restrictions, we stopped after one region. Although it could be interesting to

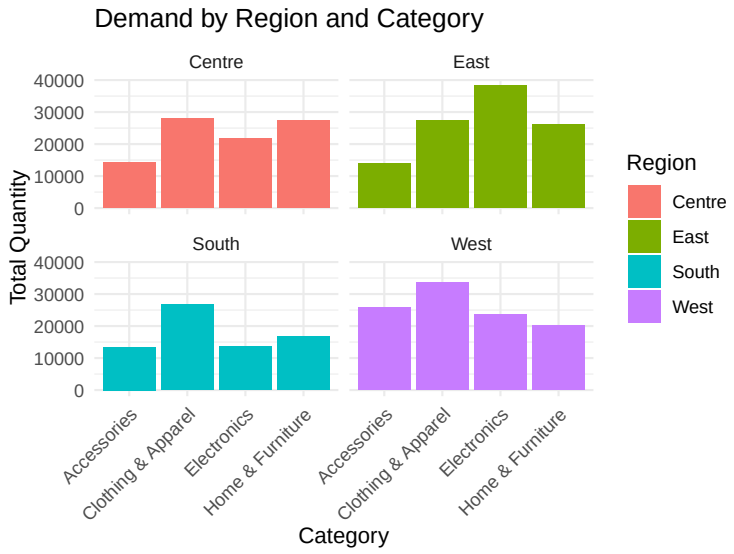

Figure 3.3: *Total Purchases per Category by Region. The distribution of demand per category varies across regions.*

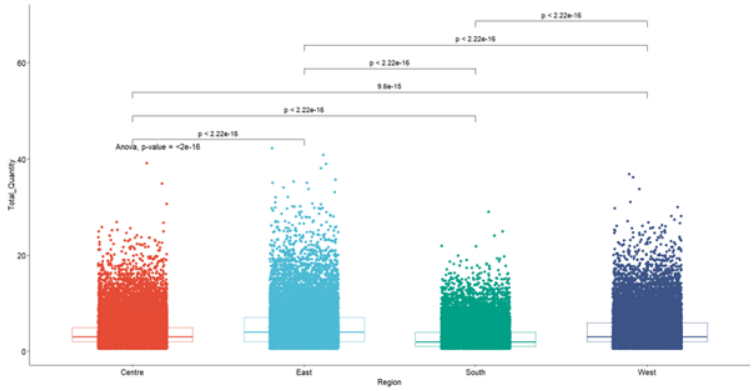

Figure 3.4: *ANOVA test results for equal mean of purchased quantities across regions. Including pairwise comparisons p-values. All resulting in highly significant differences.*

complete the analysis for each region, with this being a simulated dataset, we decided that the emphasis of the project should be on the comparison between modeling strategies rather than on the obtained results.

3.2.2 Exploring Temporal Variability

To explore demand's temporal patterns, we started by investigating seasonality. Seasonality is the presence of a periodic pattern on observations measured across time (Pennsylvania

State University, nd). Product demand tends to follow a temporal trend usually influenced by the seasons, such as holiday shopping in November and December, or summer sales around memorial and labor days. The dataset under analysis was not the exception. We can see the increase on purchases in later months in the time series plots shown in Figure 3.5

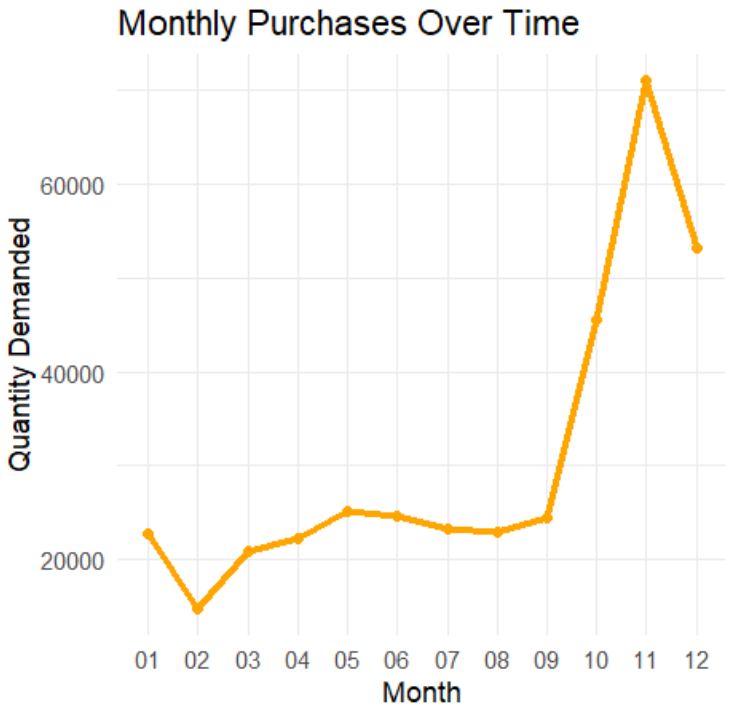

Figure 3.5: *A time series plot showing the seasonality of monthly quantity demand for the West region.*

The first method we applied to explore seasonality trends to uncover high and low demand months was principal component analysis (PCA). The principal components were generated using `princomp()` function in R that uses the correlation matrix by default, i.e., it runs the analysis in the standardized data to avoid favoring variables with larger variance. The function was applied to a transformed dataset in which every row represented a product, while each column corresponded to monthly demand values from 1 to 12. This structure was intended to evaluate if PCA would capture variation in demand patterns across each month, and whether the demand variation will indicate seasonality. However, PCA leaded to the opposite result: PC1 explained 93.6% of the total variance with all months having nearly uniform loadings (ranging from 0.28 to 0.29). This implies that products mostly differ

in overall quantity rather than distinct seasonal effects, and that seasonal variability might differ from product to product instead of being the same for all. Additionally, the PCA biplot in Figure 3.6 illustrates the loading vectors clustered tightly, reinforcing the conclusion about demand level rather than a common seasonal variability sourcing. While PC2 showed some variability in the loadings, it accounted for less than 2% of the total variance and was not considered substantially meaningful.

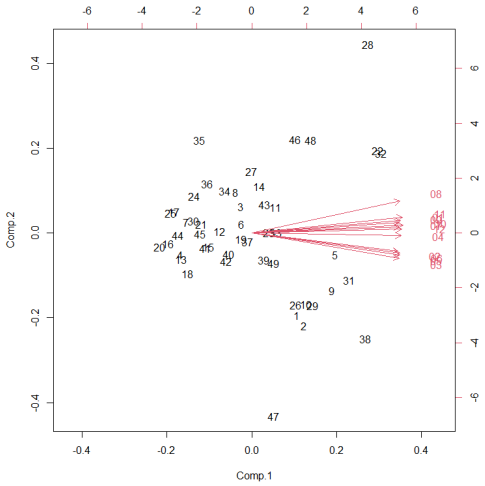

Figure 3.6: *PCA biplot of product level monthly demand, showing the projection of products onto PC1 and PC2 along with corresponding monthly loading vectors.*

The PCA results also demonstrated that a generic seasonal effect for all products would probably be unsuccessful, and that we needed to model seasonality patterns at more granular level.

The rest of the chapter will describe modeling demand at the product level, starting with predictive models and then switching to time series.

3.3 Count-Based Regression Models

Following the significant variability in demand across regions and over time revealed in the exploratory data analysis, a count based regression framework was adopted to model expected product demand on data aggregated by quarters on a city level. Due to the non-negative

integer nature of demand, the Poisson model was initially considered as a baseline model. However, exploratory analysis and statistical testing indicated the presence of overdispersion, where the variance exceeds the mean (Serra, nd). We tested overdispersion with the Cameron and Trivedi (1990) regression based test, implemented via the `AER::dispersiontest()` function in R. Under a null hypothesis of equidispersion in the Poisson model, the resulting p-value of less than 2.2×10^{-16}, provides strong evidence against the Poisson equidispersion assumption in favor of data overdispersion. As a result, a Negative Binomial regression model was employed as a more appropriate alternative to accommodate a dispersion parameter.

Consistent with the results of the overdispersion test, the Negative Binomial model demonstrated a substantially better fit compared to the Poisson model with an AIC of 23,030 which was significantly lower than the Poisson model AIC of 26,900. Out-of-sample evaluation showed that the Negative Binomial model outperformed the Poisson model at each evaluation metric. The results are summarized in Table 3.2.

Table 3.2: *Model Performance Across Methods to Predict quarterly unit purchases by product and city. Metrics were calculated using 25% observation set aside as out-of-sample data.*

Model	RMSE	MAE	Norm MAE
Poisson Regression	30.36	20.41	0.8852
Negative Binomial	10.03	6.998	0.3035

The normalized MAE (Eq. 2.3) indicated that the average prediction error of the Poisson model was approximately 86% of the mean demand. This was a very low performance compared to the Negative Binomial model's 30% normalized MAE score, which suggested a relatively strong performance.

3.4 Ordinal Regression Model

Following the results of the counts prediction, we decided to explore the prediction of demand categories rather than exact values. The product and State level dataset was divided into four classes as mentioned in Section 3.1 to maintain balanced classes for subsequent classification tasks as recommended by Sydow (2020). A Proportional Odds Logistic Regression

(POLR) model was fit to account for the ordinal nature of the demand categories. The response variable, Demand_Category was treated as an ordered factor, with State, quarter, and Product_Name included as predictors. The model performance was evaluated using a confusion matrix generated with R's `predict()` function. The model's performance resulted in an accuracy between 57.68% and 63.48% at the 95% confidence interval level, indicating moderate predictive performance.

Class level performance revealed that the model performed substantially better for extreme demand categories. Accounting for our model's sensitivity and specificity, the balanced accuracy results for High Orders (85.50%) and Very Low Orders (79.23%) were high. In contrast, the balanced accuracy results of Medium Orders (68.35%) and Low Orders (61.23%) were considerably lower confirming that the model struggled to distinguish the classes between intermediate demand levels. This pattern is consistent with common challenges in ordinal classification, where adjacent categories exhibit overlapping characteristics, making them more difficult to distinguish than more extreme categories (Pitawela et al., 2026). When analyzing the confusion matrix in Table 3.3, in light of the cost of misclassification, the model did not predict any High demand orders as Very Low, and it did not predict any Very Low orders as High orders. Overall, the POLR model is best suited for identifying extreme demand cases and may be particularly useful for decision making scenarios such as inventory reduction or identifying products with consistently low demand.

Table 3.3: *POLR Model Confusion Matrix.*

	Predicted			
Observed	Very Low	Low	Medium	High
Very Low	185	69	10	0
Low	70	104	69	2
Medium	18	88	167	64
High	0	6	46	224

3.5 Classification Tree Model

A classification tree model was implemented to predict demand categories using quarter, State, and Category as predictors as shown in Figure 3.7. Here, we used the training (75%) and testing (25%) sets to evaluate out of sample performance. A fixed random seed was used to ensure that the results were reproducible. Class proportions were consistent across both sets, ensuring balanced evaluation framework. The decision tree was trained using R's `CART::rpart()` algorithm with constraints on maximum depth and minimum bucket size to reduce overfitting. This model achieved an overall accuracy of 53.39%. This indicates that the model seems to correctly classify demand categories slightly more than half the time. Given the multi class nature of the problem, this represents moderate predictive performance, though lower than the ordinal regression model. Performance varied significantly across demand categories. The model performed relatively well for High Orders and Very Low Orders having a fairly high balanced accuracy of 72.33% and 69.67% respectively. However, the balanced accuracy for Medium Orders was poor (41.82%) and the model failed to predict any Low Orders. This suggests that the model may be biased toward more extreme categories and that the "Low Orders" class has observations that overlap significantly with adjacent categories. When looking at the confusion matrix in Table 3.4, we can see that most Low orders were predicted as Very Low and Medium with few as High orders. Overall, the decision tree model performs fairly well for extreme demand categories (high and very low orders) but struggled with the intermediate categories, particularly failing to predict Low Orders. This pattern is consistent with the challenges observed in the ordinal regression model. The absence of predictions for one class points out a huge limitation in the model's ability to distinguish classes using the current feature set and model constraints. Incorporation of additional features or auxiliary data that capture the underlying variation in demand categories may improve the model's ability to distinguish between intermediate groups.

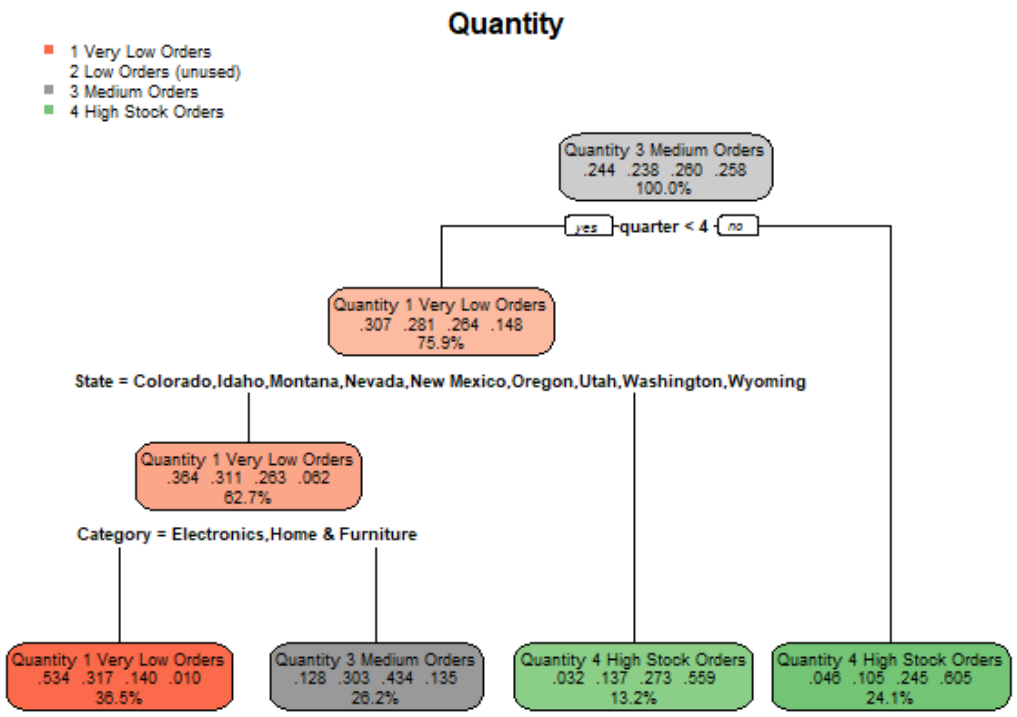

Figure 3.7: *Classification Tree used to predict demand categories illustrating how splits on quarter, state, and category partition the data into demand levels.*

Table 3.4: *Classification Tree Model Confusion Matrix.*

	Predicted			
Observed	Very Low	Low	Medium	High
Very Low	227	0	30	16
Low	118	0	90	59
Medium	58	0	118	116
High	1	0	35	254

3.6 XGBoost Model

An Extreme Gradient Boosting (XGBoost) model was implemented to classify demand categories using quarter, State, and product category as predictors. Categorical variables were transformed using one-hot encoding to create a numerical feature matrix suitable for model training. Analysis was conducted using the previously created training (75%) and testing (25%) sets, and the model was trained using a multi-class classification objective

(`multi:softprob`), which outputs class probabilities. A fixed random seed was used to ensure that the results were reproducible. The model performance was evaluated using accuracy, balanced accuracy, and a confusion matrix. The XGBoost model achieved an overall accuracy between 25.81% and 31.12% at a 95% confidence interval. This classification accuracy is very poor and close to random guessing for a four-class problem, indicating that the model was largely unable to correctly distinguish between demand categories. This can also be seen by observing the confusion matrix in Table 3.5 where the values are randomly spread across all categories. The balanced accuracy, however, had fairly weak uniform results for Very Low Orders (53.86%), Low Orders (50.91%), Medium Orders (50.70%) and High Orders (53.70%). Although performance across classes was relatively uniform, it remained weak and did not meaningfully outperform other models. It is important to take into account the cost of misclassification when comparing these models. While XGBoost had a performance comparable to other models in the intermediate categories, it misclassified a large amount of High orders in the Very Low orders category and several Very Low orders in the high orders category. One possible contributing factor to the weak model performance is the use of one-hot encoding, which substantially increased the dimensionality of the data set while introducing sparsity. Because there were many categories in the variables, the predictive signal was diluted across many sparse features (GeeksforGeeks, 2024). This made it more difficult for the model to learn meaningful and stable patterns increasing the risk of overfitting. Despite the complexity of the XGBoost algorithm, the model performed significantly worse than both the decision tree and ordinal regression models in overall accuracy. This indicates that increasing model complexity does not guarantee improved overall prediction performance. However, it improved performance prediction for some of the intermediate classes.

The results of the classification models tested in this study are summarized in Table 3.6 below. Overall the analysis highlights important trade-offs between model complexity and predictive performance. The POLR model achieved the highest overall accuracy and performed particularly well in identifying extreme demand categories, making it well suited for situations where correctly identifying high and very low demand is critical. The classifica-

Table 3.5: *XGBoost Model Confusion Matrix.*

	Predicted			
Observed	Very Low	Low	Medium	High
Very Low	77	67	72	67
Low	74	58	75	71
Medium	77	71	77	71
High	68	62	79	68

tion tree demonstrated moderate performance but failed to capture Low Orders, indicating limitations in differentiating overlapping categories. The XGBoost model produced more uniform predictions across classes, but performed poorly overall with accuracy near the random baseline. The increased model complexity did not necessarily improve performance in this setting.

Across all models, intermediate demand categories were consistently more difficult to classify due to overlapping characteristics in the variables. Additionally, evaluation based on confusion matrices highlighted the importance of considering the cost of misclassification. Misclassifying extreme categories carries greater practical consequences than misclassifying adjacent categories. The POLR model demonstrated the most favorable behavior with minimal misclassifications in extreme classes, making it the most reliable model in this context.

Table 3.6: *Comparison of Classification Model Performance on the Testing Dataset.*

Metric	**POLR**	**Decision Tree**	**XGBoost**
Overall Accuracy (%)	57.68–63.48	53.39	25.81–31.12
Balanced Accuracy by Class (%)			
High Orders	85.50	72.33	53.70
Medium Orders	68.35	41.82	50.70
Low Orders	61.23	0.00	50.91
Very Low Orders	79.23	69.67	53.86

3.7 Time Series Modeling

Time series models were developed to analyze and forecast potential trends in individual product demand over time. For each product in the dataset, a time series was constructed using the sales observations aggregated by month and quarter respectively across the entire

Region. For the monthly time series (Monthly Model), a 24 months horizon was fitted for each product while for the quarterly time series (Quarterly Model), an 8 quarters horizon was fitted for each product. Both models were fitted using ARIMA models at a 95% confidence level. Decomposition techniques were also used to examine the trend and seasonal components of the series. Forecast accuracy was evaluated using standard time series performance metrics, like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE) as defined in Section 2.6.

3.7.1 Monthly Model Results

The results from this monthly time series model exhibited improved forecasting performance compared to the quarterly models, although overall accuracy remained limited. A summary table of the evaluation metrics for each product is shown in Table 3.7. The mean RMSE of 33.70 indicated large forecasting error with forecasts deviating by approximately 34 units on average. Similarly, the MAE of 25.46 suggested large deviations in the forecasts of about 25 units per period. The average Mean Absolute Scaled Error (MASE) was 1.63. This is greater than 1, indicating significantly worse performance than a naive forecasting approach. This suggests that the models were unable to capture the high variability in the monthly data.

3.7.2 Quarterly Model Results

Contrastingly, the quarterly ARIMA models exhibited substantially worse forecasting performance across most products. The summary table of the evaluation metrics for each product is shown in Table 3.7. The mean RMSE of 81.06 indicates large forecasting errors are significantly higher than those observed in the monthly models. Similarly, the MAE of 66.99 suggests that forecasts deviated by approximately 67 units per period. The Mean Absolute Scaled Error (MASE) values had an average of 2.88, which is well above 1, indicating that the models performed worse than a naive forecasting approach. Additionally, the high

variability in MASE values suggests inconsistent model performance across products.

Table 3.7: *Summary of Monthly and Quarterly Model Evaluation Metrics Across Products*

	Monthly			Quarterly		
Statistic	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
Min	11.36	8.97	0.8692	11.36	8.97	0.8692
1st Qu.	20.72	16.29	1.2135	32.17	23.77	1.45
Median	32.50	23.80	1.4784	55.17	43.50	2.0837
Mean	33.70	25.46	1.5827	81.06	66.99	2.8826
3rd Qu.	42.94	32.96	1.7654	127.75	107.38	3.3332
Max	74.47	55.34	2.8912	280.61	235.16	26.8300

In summary, the monthly models outperformed the quarterly models across all evaluation metrics. While the MASE values above 1 indicate that both approaches struggled with forecasting accuracy, the quarterly aggregation resulted in a significant loss of information, leading to poorer model performance. The aggregation of the data for the quarterly models reduced the number of observations available for model fitting, limiting the ability of ARIMA models to accurately estimate temporal dependencies. Additionally, important variability present in the monthly data was lost during aggregation, leading to poorer predictive performance.

3.8 Conclusion

In conclusion, the results demonstrate that model performance varied substantially depending on both the structure of the data and the modeling approach. Among the regression approaches, the analysis conducted on a city and product level dataset produced results summarized in Table 3.2. These results highlighted that the Negative Binomial model provided the best fit for count-based demand data. By accounting for overdispersion, it achieved stronger predictive performance compared to the Poisson model.

For classification tasks, also performed on a city and product level dataset, the comparison of models were summarized in Table 3.6. These results revealed important trade offs between overall accuracy and class level performance. The ordinal regression and decision tree models achieved moderate overall accuracy, performing well in identifying extreme

demand categories but struggled with the intermediate classes, likely due to overlapping characteristics between adjacent categories. In contrast, the XGBoost model had uniform but weak performance across all classes. It exhibited substantially lower overall accuracy indicating poor predictive reliability. The model suffered from reduced accuracy likely due to feature limitations and the effects of high-dimensional sparse representations introduced through one-hot encoding. Additionally, the analysis of the confusion matrices demonstrated that the POLR model had the most reliable behavior by minimizing severe misclassifications between extremely different categories whereas the XGBoost model had a high frequency of these potentially more serious errors. For an ordinal response such as demand-level, classification methods should control for more "costly" misclassifications into classes further apart in the ordinal sequence. Overall, the POLR model provided the best performance when considering both accuracy and the potential cost of misclassification.

Time series analysis performed on a product level across the region dataset produced results that were summarized in Table 3.7. They showed that monthly models consistently outperformed quarterly models across all evaluation metrics. While both approaches exhibited forecasting challenges, the quarterly aggregation led to a significant increase in error, with ARIMA models performing worse than naive benchmark. In contrast, the monthly models retained more underlying variability and produced relatively more reliable forecasts. These findings suggest that given the dataset only had two years, aggregating to a lower frequency resulted in loss of important temporal information, and reduced the model effectiveness.

Overall, the results suggest that model performance depends on appropriate data representation, distributional assumptions, and alignment with the problem structure rather than on algorithmic complexity alone.

Chapter 4

Discussion and Future Work

The analyses conducted in this study reveal meaningful structure in quarterly and monthly product-level demand and demonstrate the value of combining statistical and machine learning models to understand retail dynamics. The variability observed across products, states, and months highlights the importance of using modeling approaches that can accommodate nonlinearity and product specific behavior.

In the count-based regression analysis, identifying overdispersion in the data played an important role in model selection. The Poisson model, which assumes equality of the mean and variance, produced poorer results, due to its inability to account for excess variability in demand. Contrastingly, the Negative Binomial model provided improved predictive performance.

The classification models showed strong performance when predicting high demand and very low demand observations, suggesting their usefulness in identifying extreme demand scenarios for inventory planning. However, performance was consistently weaker when predicting demand for medium and low demand categories, indicating that these cases are inherently more difficult to distinguish. Incorporating auxiliary data with additional explanatory variables may help capture the variability needed to improve classification performance in these categories.

From a time series perspective, the monthly ARIMA models outperformed the quarterly

models across all evaluation metrics. Although forecasting accuracy remained limited overall, the higher frequency monthly data retained important variability that was lost during temporal aggregation in the quarterly model.The quarterly models exhibited substantially higher error, likely due to the reduced number of observations and loss of information in aggregation. These results suggest that in this context, aggregating data to lower temporal resolution can degrade model performance. While aggregation is often used to reduce noise, it may also eliminate meaningful patterns that are necessary for accurate forecasting. While this emphasizes the importance of selecting an appropriate time scale to balance reducing noise with signal preservation, it also indicated a need for more time periods for improved time series model performance.

4.1 Limitations

A key limitation of this study is the use of synthesized data, which may not fully capture the complexity and variability of real-world retail environments. Additionally, the absence of a unique customer identifier restricted the ability to analyze customer-level behavior, such as repeat purchasing patterns or customer segmentation.

Another challenge was that some products were aggregated by a generic name without differentiation by size, type, or unique product identifier. This led to extreme variation in price for some products. While price was reintroduced in the aggregated data as mean price in an attempt to capture pricing effects, it was found not statistically significant. It appears that mean price became a proxy for the product name and the strong association between the mean price, product name and category confounded the effect of price on demand.

Model assumptions also present potential limitations. For example, time series models assume a certain degree of stationarity, and count models rely on distributional assumptions that may not fully hold in practice. Violations of these assumptions may impact model performance.

Finally, challenges related to source reliability and validation highlight the importance of carefully verifying references and ensuring the credibility of supporting literature when

conducting research in data analysis.

4.2 Future Work

Future work could focus on incorporating more granular auxiliary data, including unique customer-level identifiers, additional customer information, promotional variables, and detailed pricing and product information. Improving the representation of price, such as incorporating variability or volatility as a pricing measure, could enhance model performance.

Additionally, further investigation into simulation based data generation methods could improve the realism and usability of synthetic datasets for both learning purposes and model development. This would allow for more robust evaluation and improved applicability to real world settings.

Lastly, since the dataset included only a limited number of cities per state, spatial interpolation methods such as Kriging could be used to estimate demand patterns in areas without existing store locations. This approach would support more informed decisions regarding market expansion and store placement.

Bibliography

Amperon (2023). Understanding the benefits of nmae over mape for estimating load forecast accuracy. Accessed: March 29, 2026.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Routledge.

Calzone, O. (2022). Mae, mse, rmse, and f1 score in time series forecasting. Accessed: March 29, 2026.

Cameron, A. C. and P. K. Trivedi (1990). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics 46*(3), 347–364.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 785–794. Association for Computing Machinery.

Gates, S. (2025). What is data completeness? definition, examples, and kpis. Accessed: 2026-03-29.

GeeksforGeeks (2024). One hot encoding in machine learning. Accessed: 2026-03-29.

Google Developers (2024). Accuracy, precision, recall. Accessed: 2026-03-29.

Grandini, M., E. Bagli, and G. Visani (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*.

Hickins, M. (2023). What is retail analytics? https://www.oracle.com/retail/what-is-retail-analytics/. Oracle Corporation. Accessed: 2026-04-02.

Hyndman, R. and G. Athanasopoulos (2018). *Forecasting: principles and practice, 2nd edition*. OTexts: Melbourne, Australia. Accessed on 3/21/26.

Hyndman, R. J. and G. Athanasopoulos (2021). Forecast accuracy measures. Accessed: March 29, 2026.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2 ed.). New York: Springer.

O'Brien, K. (2022). Akaike information criterion. RPubs article. Accessed: March 29, 2026.

Pennsylvania State University (n.d.). Stat 510: Applied time series analysis – lesson 4. https://online.stat.psu.edu/stat510/Lesson04. Accessed: 2026-04-02.

Pitawela, D., G. Carneiro, and H.-T. Chen (2026). Cloc: Contrastive learning for ordinal classification with multi-margin n-pair loss.

Serra, I. (n.d.). Testing overdispersion. https://rpubs.com/DonArres/TestingOverdispersion. RPubs article, accessed March 23, 2026.

Sydow, D. (2020). The importance of a balanced dataset in machine learning classification algorithms. https://medium.com/@dsydow/the-importance-of-a-balanced-dataset-in-ml-classification-algorithms-3a55dff768f7. Accessed: 2026-03-31.

Tredence (2024). Retail demand forecasting in 2025: Types, importance. Accessed: 2026-03-30.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4 ed.). New York: Springer. Accessed via PDF: https://www.mimuw.edu.pl/~pokar/StatystykaMgr/Books/VenablesRipley_ModernAppliedStatisticsS02.pdf.

XGBoosting (2023). Configure xgboost "multi:softprob" objective. Accessed: March 29, 2026.

Yennewar, Y. (2023). Product sales dataset 2023–2024. https://www.kaggle.com/datasets/yashyennewar/product-sales-dataset-2023-2024. Accessed: 2026-04-02.