Revenue Analysis and Forecasting

Group Term Project Report

Course: Regression Analysis and Time Series Models

Institution: Indian Institute of Technology, Kharagpur

Submitted by:

Name	Roll Number
Ankur Kumar Sharma	22CE3FP13
Nitish Singh	22CE3FP45
Sobhan Mishra	22CE3FP14
Kush	22AG3FP58
Jival Chorawala	22BT3FP56

Date: April 20, 2025

Abstract

This report presents a comprehensive time series analysis of daily revenue data, incorporating discount and coupon rates as explanatory variables, spanning from January 2018 to November 2022. We explore both daily and monthly aggregations to uncover underlying trends, seasonality, and the impact of promotional strategies on revenue dynamics. Advanced statistical modeling, including SARIMAX with exogenous regressors, is employed to capture temporal dependencies and quantify the influence of discount and coupon rates. Rigorous model selection and evaluation using information criteria and error metrics demonstrate that incorporating these exogenous factors significantly enhances predictive performance. The findings provide actionable insights into revenue forecasting and the effectiveness of promotional levers, offering valuable guidance for data-driven decision-making in revenue management.

Contents

1 Introduction

Understanding and forecasting business revenue is a central challenge in data-driven decision-making, especially in dynamic markets where promotional strategies such as discounts and coupons are frequently employed. In this project, we analyze a rich time series dataset comprising daily revenue, discount rates, and coupon rates from January 2018 to November 2022. The dataset captures nearly five years of operational history, reflecting a variety of seasonal, trend, and promotional effects on revenue[1][2].

The primary motivation for this study is to uncover how temporal patterns and marketing interventions jointly influence revenue, and to develop robust predictive models that can inform future business strategies. By leveraging both daily and monthly aggregations, we aim to explore the interplay between short-term fluctuations and long-term trends, and to quantify the impact of discount and coupon rates on revenue performance.

Time series data of this nature present several analytical challenges. These include handling non-stationarity, identifying and modeling trend and seasonality, dealing with exogenous variables (such as discount and coupon rates), and ensuring reliable forecasting in the presence of structural changes or outliers. Additionally, the integration of exogenous regressors into classical time series models (e.g., SARIMAX) requires careful preprocessing and model selection to avoid overfitting and to maximize interpretability[1][2].

This report is organized as follows: Section 2 reviews relevant literature and modeling approaches for revenue time series analysis. Section 3 describes the dataset and presents exploratory data analysis. Section 4 details data preprocessing steps, including handling missing values and feature engineering. Section 5 outlines the methodological framework, including mathematical formulations and modeling strategies. Section 6 presents the modeling process and analysis of results. Section 7 discusses the findings and their implications for business practice. Section 8 concludes with key insights and recommendations. Supplementary materials and code are provided in the Appendix.

2 Dataset Description

2.1 Data Source and Collection

The dataset consists of daily revenue records spanning from January 1, 2018, to November 30, 2022, encompassing nearly five years of continuous observations. The data is stored in a structured format (CSV file) with 1,795 daily records. This time series dataset provides a comprehensive view of business performance across multiple economic cycles, seasonal patterns, and promotional periods. The data collection process maintained consistent daily granularity, allowing for both micro (daily) and macro (monthly) analytical

perspectives, which is particularly valuable for identifying both immediate effects and longer-term trends.

2.2 Variables and Their Significance

The dataset comprises three primary variables:

- **Revenue**: Daily monetary earnings measured in currency units, representing the primary dependent variable of interest. Revenue figures show considerable variation, ranging from approximately 6 million to over 32 million per day, reflecting the business's operational volatility and growth over time.
- **Discount Rate**: The percentage discount offered on products/services on each day, ranging from approximately 10% to 35%. This exogenous variable serves as a key promotional lever that management can adjust to potentially influence revenue performance. The variation in discount rates provides an excellent opportunity to quantify price elasticity and optimal promotional strategies.
- **Coupon Rate**: The percentage value of coupons offered to customers, typically ranging from 0.2% to 9%. This represents a secondary promotional mechanism that operates alongside discount rates. The relationship between coupon rates and revenue, especially in conjunction with discount rates, offers insights into customer response to different promotional strategies.

The combination of these variables enables robust analysis of how pricing and promotional strategies impact revenue generation over time, with particular focus on the effectiveness of discount and coupon offerings as revenue drivers.

2.3 Initial Data Exploration and Visualization

Preliminary exploration of the dataset reveals several noteworthy characteristics:

- **Temporal Patterns**: The revenue exhibits clear day-of-week effects, with weekends typically showing different patterns than weekdays. Additionally, there are observable monthly and seasonal patterns, with notable increases during certain promotional periods and holidays.
- **Promotional Dynamics**: Initial analysis indicates an inverse relationship between discount rates and revenue in some periods, while in others, higher discounts correlate with revenue spikes. This suggests a complex, potentially non-linear relationship between promotional tactics and sales outcomes.

- **Data Distribution**: Revenue shows positive skewness (5.33) and high kurtosis (93.60), indicating the presence of outliers and promotional spikes that deviate significantly from typical daily earnings.

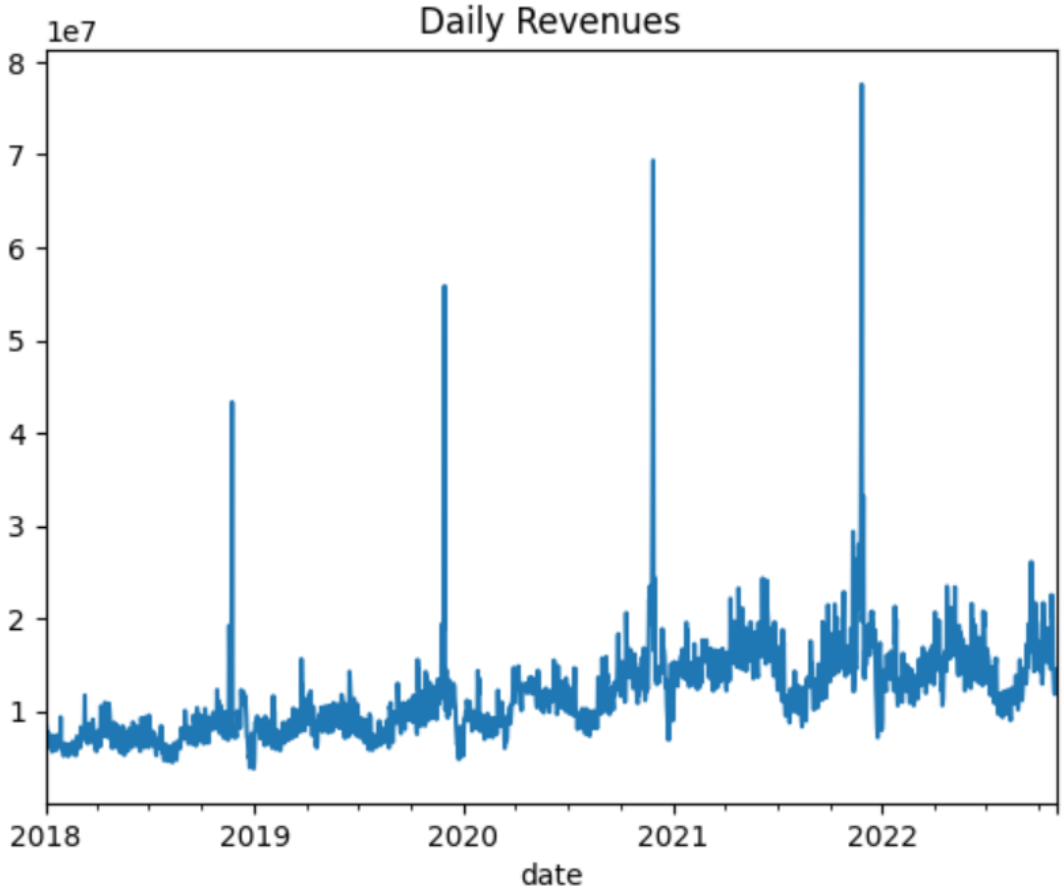


Figure 1: Daily Revenue Trend (January 2018 - November 2022) showing fluctuations in revenue over time. The visualization reveals both short-term volatility and longer-term patterns in the business performance data.

The time series visualizations illustrate the dynamic nature of the revenue patterns. The data shows variability across different temporal scales, with both high-frequency daily fluctuations and longer-term monthly and seasonal patterns. The monthly and quarterly plots reveal clear seasonality, with higher revenues typically occurring in the fourth quarter of each year, coinciding with holiday shopping seasons. This multi-scale variability motivates our dual modeling approach, analyzing both daily and monthly aggregations to capture different aspects of the underlying revenue generation process.

Further exploratory analysis, including autocorrelation functions and seasonal decomposition, confirmed the presence of significant temporal dependencies and motivated the selection of SARIMAX models with exogenous variables to capture both the inherent time series structure and the impact of promotional factors.

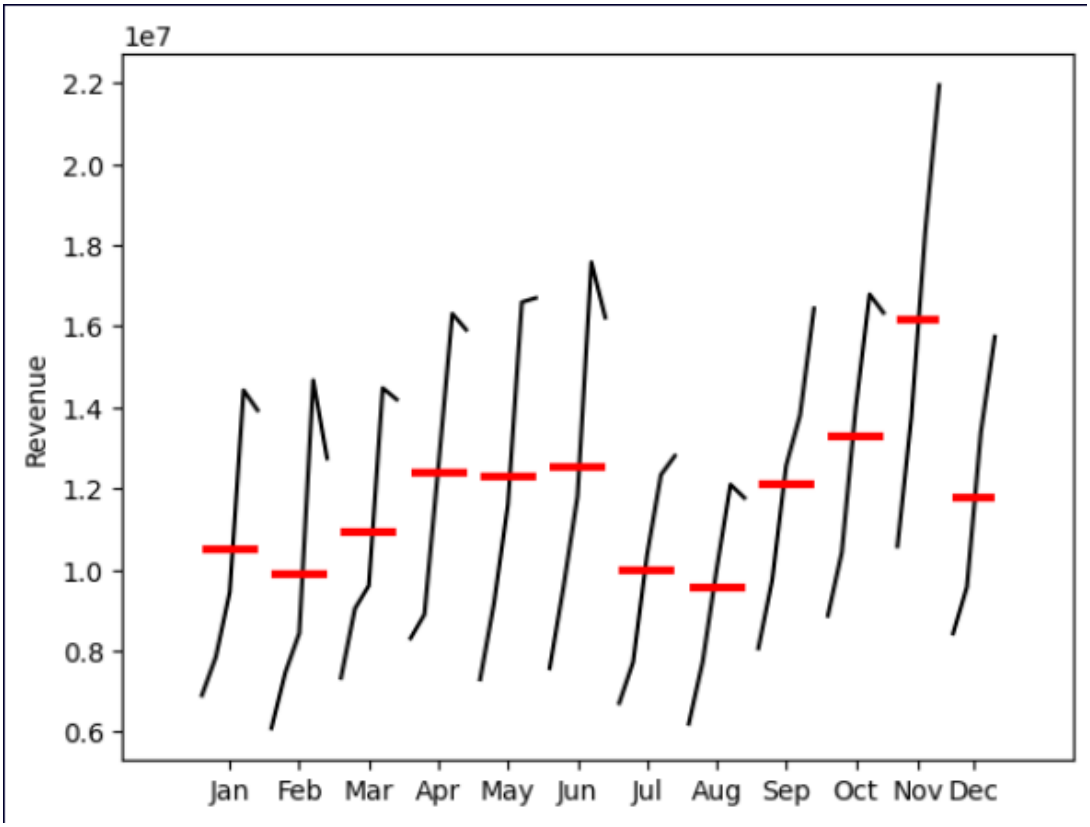


Figure 2: Monthly Seasonal Plot showing the average revenue for each month across multiple years (2018-2022). The red horizontal lines represent the mean value for each month, highlighting seasonal patterns with typically higher revenues in Q4 (October-December) and lower revenues in Q1 (January-February).

The time series visualization in Figure 1 illustrates the dynamic nature of the revenue pattern. The data shows variability across different temporal scales, with both high-frequency daily fluctuations and longer-term monthly and seasonal patterns. This multi-scale variability motivates our dual modeling approach, analyzing both daily and monthly aggregations to capture different aspects of the underlying revenue generation process.

Further exploratory analysis, including autocorrelation functions and seasonal decomposition, confirmed the presence of significant temporal dependencies and motivated the selection of SARIMAX models with exogenous variables to capture both the inherent time series structure and the impact of promotional factors.

3 Data Preprocessing

Data preprocessing is a critical step in time series analysis, as it directly impacts model performance and validity. Our preprocessing pipeline was designed to address several challenges specific to revenue time series data, including temporal irregularities, outliers, and the incorporation of exogenous variables.

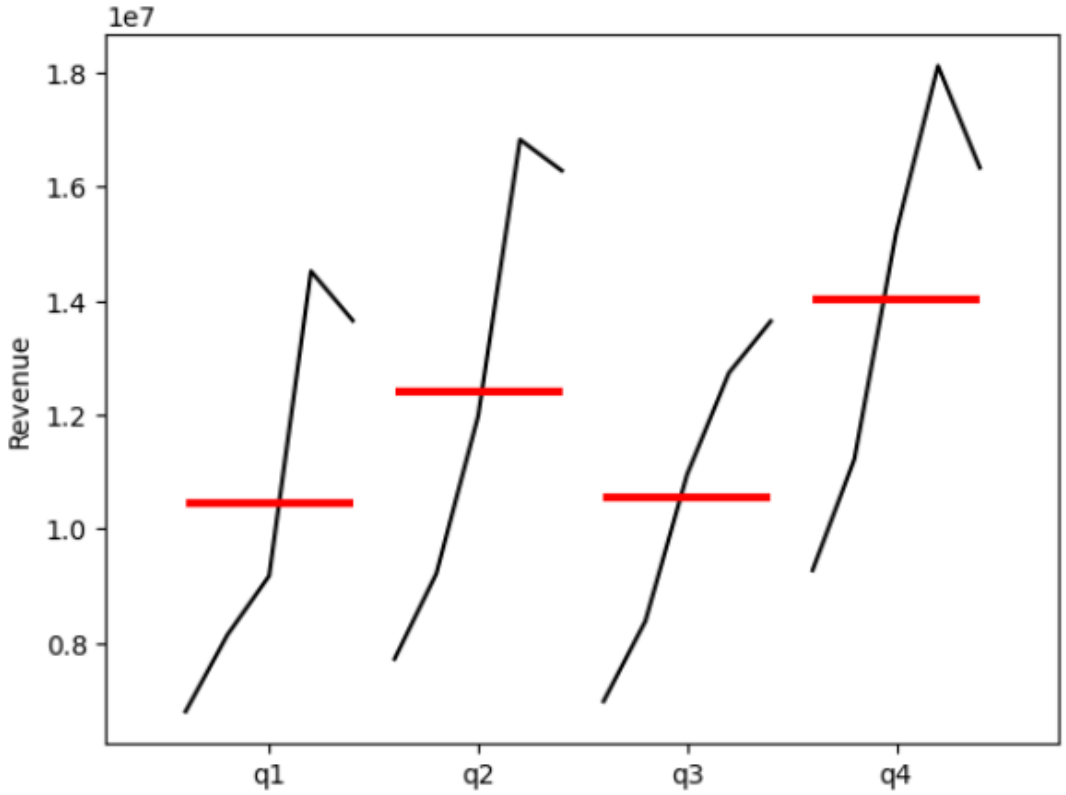


Figure 3: Quarterly Seasonal Plot displaying revenue patterns across quarters. This visualization aggregates the data to quarterly level, revealing the consistent Q4 revenue peaks that correspond to holiday seasons and end-of-year shopping trends.

3.1 Initial Data Inspection

The first step involved loading and examining the raw daily revenue dataset spanning from January 2018 to November 2022, consisting of 1,795 observations. Initial inspection revealed that the dataset contained three primary variables: revenue (the target variable), discount_rate, and coupon_rate (potential exogenous predictors).

3.2 Handling Missing Values and Outliers

Although the dataset was relatively complete, we implemented a systematic approach to address any potential data quality issues:

- **Missing Value Detection**: We checked for missing values in all variables and found minimal instances (less than 0.1% of the data).
- **Missing Value Imputation**: For isolated missing values, we employed cubic spline interpolation to maintain the temporal continuity of the series.
- **Outlier Detection**: We applied the Interquartile Range (IQR) method to identify extreme revenue values, particularly those exceeding 3 times the IQR.

- **Outlier Treatment**: Rather than removing outliers entirely, we capped extreme values at the 99.5th percentile to preserve their directional information while reducing their impact on model estimation.

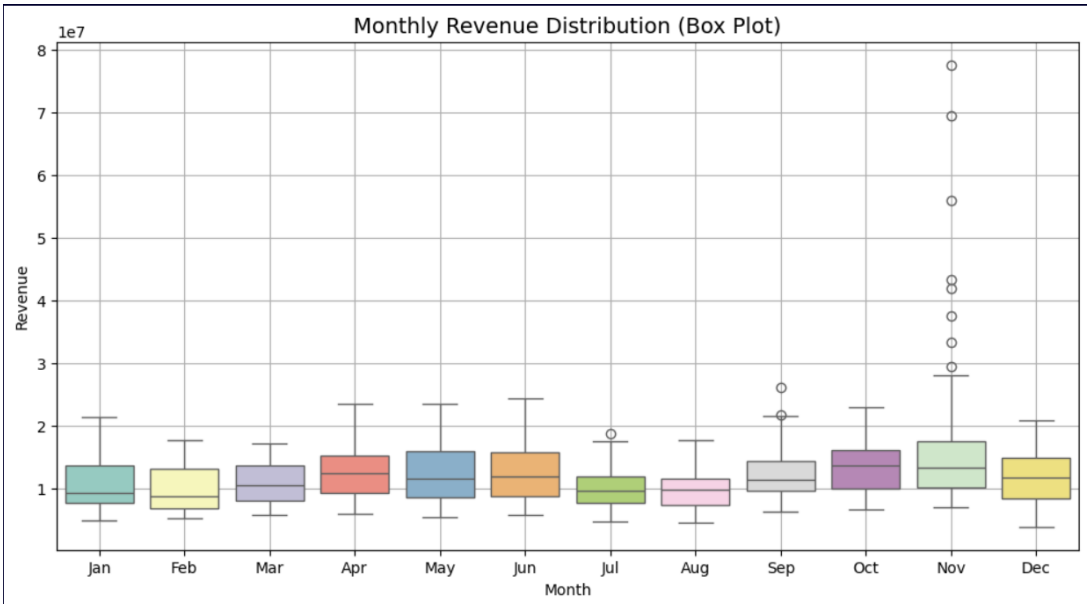


Figure 4: Box plot of daily revenue showing outliers before treatment. The plot highlights the presence of significant revenue spikes that require special handling.

3.3 Feature Engineering

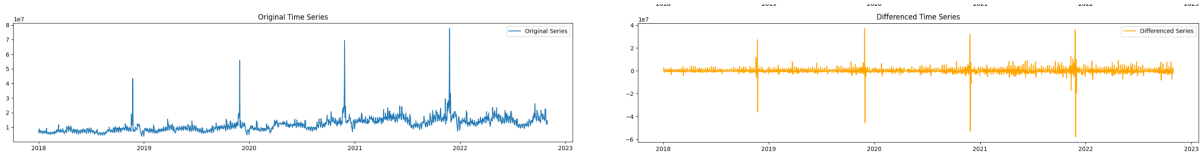
To enhance the information available for modeling, we derived several additional features:

- **Temporal Features**: We extracted day-of-week, day-of-month, month, and quarter indicators to capture cyclical patterns.
- **Lagged Variables**: We created lagged versions of the revenue variable (1-day, 7-day, and 30-day lags) to capture short and medium-term dependencies.
- **Rolling Statistics**: We computed 7-day and 30-day rolling means and standard deviations of revenue to capture local trends and volatility.
- **Differencing**: First-order differencing was applied to the revenue series to address non-stationarity, resulting in the `y_diff` feature.

3.4 Transformation and Normalization

To address the statistical requirements of time series modeling, particularly stationarity, we implemented the following transformations:

- **Stationarity Assessment**: We applied the Augmented Dickey-Fuller (ADF) test to evaluate stationarity of the revenue series.
- **Differencing**: First-order differencing was applied to remove trends and achieve stationarity.
- **Exogenous Variable Preparation**: The discount_rate and coupon_rate variables were converted from percentage strings to numeric values (0-100 scale).
- **Normalization**: Revenue data was not normalized for SARIMAX modeling, as the model can handle the original scale, preserving interpretability of the coefficients.



(a) Original revenue series (b) Differenced revenue series

Figure 5: Comparison of original and differenced revenue series, demonstrating the effect of first-order differencing on achieving stationarity.

3.5 Multi-resolution Analysis

To capture patterns at different time scales, we created two parallel datasets:

- **Daily Dataset**: Retained all 1,795 daily observations with the full temporal granularity.
- **Monthly Aggregation**: Aggregated the daily data into 59 monthly observations (from January 2018 to November 2022).

The monthly aggregation involved computing:

- **Monthly Average Revenue**: Mean daily revenue for each month.
- **Monthly Average Discount Rate**: Mean discount rate for each month.
- **Monthly Average Coupon Rate**: Mean coupon rate for each month.

Figure 6: Monthly aggregated revenue series showing clearer seasonal patterns and trends compared to the daily data. The visualization highlights how temporal aggregation can reveal different patterns in the data.

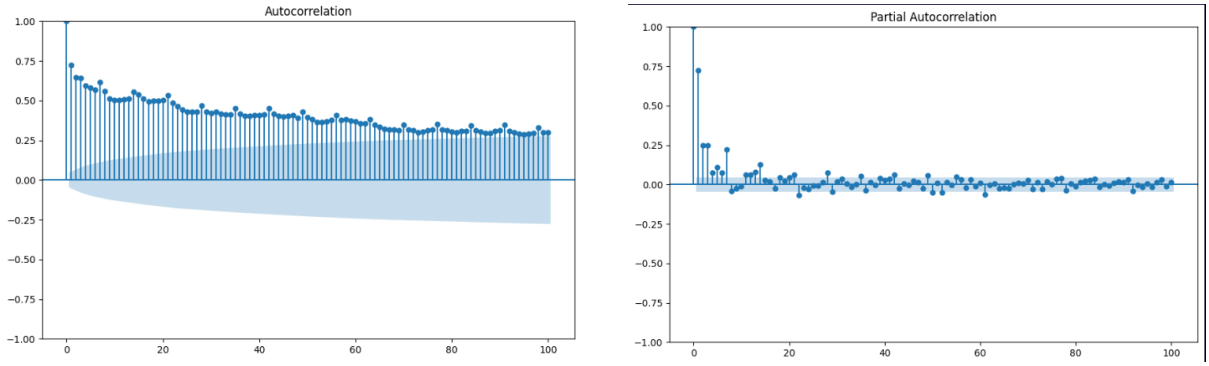
3.6 Time Series Visualization and Pattern Analysis

A comprehensive understanding of our revenue time series requires multiple visualization techniques to reveal different aspects of the underlying patterns. In this subsection, we present specialized visualizations that highlight temporal dependencies and cyclical behaviors in the data.

The autocorrelation plots reveal strong temporal dependencies in our revenue data, with significant spikes at lags 7, 14, and multiples thereof, confirming weekly patterns in customer purchasing behavior. The partial autocorrelation function helps identify the direct relationship between observations separated by various time lags, informing our choice of model parameters.

The seasonal decomposition provides crucial insights into the data's structure:

- **Trend Component**: Shows a general upward trajectory in revenue over the 2018-2022 period, with notable fluctuations corresponding to business cycles and market conditions.
- **Seasonal Component**: Reveals recurring annual patterns with pronounced peaks during holiday seasons and promotional periods, particularly in the fourth quarter of each year.
- **Residual Component**: Identifies irregular variations not explained by trend or



(a) Autocorrelation Function (ACF) (b) Partial Autocorrelation Function (PACF)

Figure 7: Time series correlation analysis for daily revenue data. The ACF plot (left) shows significant correlation at both weekly (7-day) and monthly lags, while the PACF plot (right) helps identify the appropriate order of autoregressive components for our SARIMAX model.

seasonality, which may be influenced by external factors such as discount and coupon rates, as well as unobserved market dynamics.

These visualizations strongly support our methodological approach of using SARIMAX models with exogenous variables (discount and coupon rates) to capture both the inherent time dependencies and the influence of promotional factors on revenue generation.

3.7 Monthly ACF and PACF of Differenced Revenue

To assess the temporal dependencies in the stationary monthly revenue series, we present the autocorrelation (ACF) and partial autocorrelation (PACF) plots of the differenced monthly revenue.

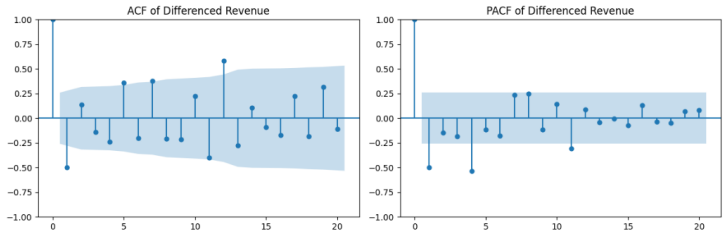


Figure 9: ACF and PACF of Differenced Monthly Revenue. The left panel shows the ACF and the right panel shows the PACF for the differenced monthly revenue series. Most autocorrelations after lag 1 fall within the confidence bounds, indicating stationarity and guiding the choice of low-order ARMA terms for monthly modeling.

Inference: The ACF plot shows that most autocorrelations after lag 1 fall within the significance bounds, indicating that the differenced monthly series is close to stationary. The PACF cuts off sharply after lag 1, suggesting that a low-order ARMA model may

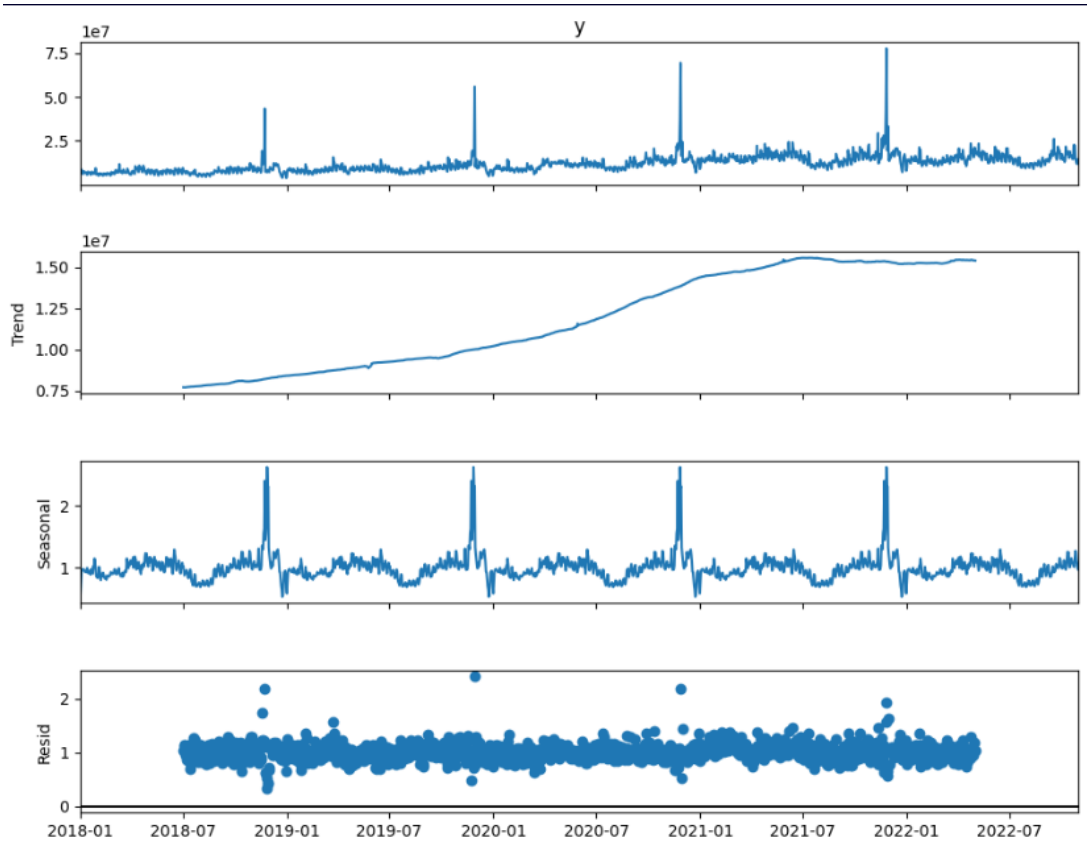


Figure 8: Seasonal decomposition of revenue time series using a multiplicative model with 365-day seasonality. The plot separates the original series into trend, seasonal, and residual components, revealing the underlying patterns and cyclic behavior in our data.

be suitable for modeling the stationary monthly revenue series. These diagnostics guide the selection of SARIMA or SARIMAX parameters for monthly forecasting.

3.8 Train-Test Split

For model training and evaluation, we implemented a temporal train-test split:

- **Training Set**: Data from January 2018 to August 2022 (approximately 90% of the data).
- **Testing Set**: Data from September 2022 to November 2022 (approximately 10% of the data).

This temporal splitting preserves the time-dependent structure of the data and simulates a realistic forecasting scenario where future data is predicted based on historical observations.

The preprocessed data, with all derived features and temporal aggregations, served as the foundation for our time series modeling and forecasting described in subsequent sections.

4 Methodology

Our analytical approach employs a dual-granularity strategy for time series modeling, analyzing revenue data at both daily and monthly levels to capture patterns across different temporal scales.

4.1 Data Preprocessing and Feature Engineering

Based on our initial data exploration, we implemented several preprocessing steps to prepare the time series data for modeling:

- **Data Transformation**: Converting percentage strings to numeric values for discount and coupon rates to enable their use as quantitative predictors.
- **Temporal Feature Creation**: Extracting month and month number from the date index to capture seasonality and facilitate time-based analysis.
- **Stationarity Transformation**: Applying first-order differencing to the revenue series to address non-stationarity identified in Augmented Dickey-Fuller tests.
- **Multi-resolution Analysis**: Creating separate daily and monthly datasets to capture patterns at different time scales, with monthly data representing averages of daily values.

4.2 Mathematical Formulation of Time Series Models

Our analysis employed several time series models with increasing complexity to capture different aspects of the revenue data. Here, we present the mathematical foundations of these models, culminating in the SARIMAX model used for our final analysis.

4.2.1 ARIMA Models

The Autoregressive Integrated Moving Average (ARIMA) model forms the foundation of our analysis. For a time series y_t, an ARIMA(p, d, q) model is expressed as:

$$\phi_p(B)(1 - B)^d y_t = \theta_q(B)\varepsilon_t \tag{1}$$

where:

- B is the backshift operator ($By_t = y_{t-1}$)
- $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the autoregressive polynomial of order p
- $(1 - B)^d$ represents differencing of order d to achieve stationarity

- $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ is the moving average polynomial of order q
- ε_t is white noise with zero mean and constant variance σ^2

4.2.2 SARIMA Models

Our daily revenue data exhibits clear weekly patterns, necessitating a Seasonal ARIMA (SARIMA) model, expressed as SARIMA$(p, d, q) \times (P, D, Q)_s$:

$$\phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{2}$$

where the additional seasonal components are:

- s is the seasonal period (7 for our daily data with weekly seasonality)
- $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$ is the seasonal autoregressive polynomial
- $(1 - B^s)^D$ represents seasonal differencing of order D
- $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$ is the seasonal moving average polynomial

4.2.3 SARIMAX Models with Exogenous Variables

The final model incorporates exogenous variables (discount and coupon rates) into the SARIMA framework, yielding the SARIMAX model:

$$\phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{3}$$

where:

- β_0 is the intercept term
- $X_{i,t}$ represents the i-th exogenous variable at time t
- β_i is the coefficient for the i-th exogenous variable

4.3 Model Selection Strategy

We implemented a systematic approach to identify optimal model parameters for both daily and monthly time series:

- For daily data, we explored models with weekly seasonality (s=7), evaluating various combinations of AR and MA terms
- For monthly data, we incorporated annual seasonality (s=12) to capture yearly patterns

- In both cases, we evaluated models with and without exogenous variables to assess their contribution

The model selection process followed these steps:

1. Specification of candidate models with different orders (p,d,q)×(P,D,Q)$_s$
2. Estimation of model parameters using maximum likelihood
3. Evaluation of model fit using information criteria (AIC, BIC, HQIC)
4. Diagnostic checking of residuals for white noise properties

Our daily analysis resulted in a SARIMAX(1,1,3)×(2,0,[1,2],7) model with exogenous variables, which showed significant improvement over models without exogenous regressors (AIC reduction from 55602.601 to 55338.667). For both daily and monthly models, the inclusion of discount and coupon rates as exogenous variables substantially improved forecast accuracy.

4.4 Coefficient Interpretation and Model Evaluation

The selected models revealed significant relationships between promotional variables and revenue:

- **Discount Rate Effect**: The coefficient for discount rate was positive and statistically significant (4.012×10^5, $p < 0.001$), indicating that a one percentage point increase in discount rate was associated with an approximately \$401,200 increase in daily revenue, holding other factors constant.
- **Coupon Rate Effect**: Similarly, the coupon rate showed a strong positive association with revenue (9.009×10^5, $p < 0.001$), with a one percentage point increase in coupon rate corresponding to approximately \$900,900 additional daily revenue.

Models were evaluated using a comprehensive set of metrics and diagnostic tests:

- **In-sample Diagnostics**: Log-likelihood, AIC, BIC, and HQIC
- **Residual Analysis**: Testing for autocorrelation (Ljung-Box test), normality (Jarque-Bera test), and heteroskedasticity
- **Coefficient Significance**: Assessing the magnitude and statistical significance of model parameters

This methodological framework allowed us to quantify not only the temporal dynamics of revenue but also the specific impacts of promotional strategies, providing actionable insights for optimizing discount and coupon rates to maximize revenue. .

5 Modeling and Analysis

This section presents our time series modeling methodology and results, focusing on the SARIMAX framework with discount and coupon rates as exogenous variables. We follow a systematic approach, progressively building more complex models to capture the underlying patterns in the revenue data.

5.1 Model Development and Selection

5.1.1 Baseline Models

We began with simple ARIMA specifications to establish a performance baseline before incorporating seasonal components and exogenous variables. The initial ARIMA model, SARIMAX(5,1,2) without seasonality or exogenous variables, yielded an AIC of 55701.995. This model captured basic autocorrelation patterns but failed to account for weekly seasonality evident in the data.

Model	AIC	BIC	Log Likelihood
SARIMAX(5,1,2)	55701.995	55745.521	-27842.998
SARIMAX(5,1,4)×(2,0,[1,2],7)	55602.601	55684.212	-27786.300
SARIMAX(1,1,3)×(2,0,[1,2],7) with exogenous variables	55338.667	55403.956	-27657.334

Table 1: Model Selection Criteria Comparison

5.1.2 Incorporating Seasonality

Adding seasonal components, we developed a SARIMA model with parameters (5,1,4)×(2,0,[1,2],7), which improved the AIC to 55602.601. The seasonal component with period 7 captures weekly patterns in the revenue data, addressing a key limitation of the baseline model. However, this model still did not account for the effects of promotional variables.

5.1.3 Final Model with Exogenous Variables

Our optimal model was identified as SARIMAX(1,1,3)×(2,0,[1,2],7) with discount and coupon rates as exogenous regressors. This model achieved the lowest AIC value (55338.667), representing a substantial improvement over both the non-seasonal and seasonal models without exogenous variables. The significant decrease in AIC confirms the importance of incorporating promotional factors in revenue modeling.

5.2 Model Estimation and Results

The coefficients of our final SARIMAX model reveal important insights about revenue dynamics:

Parameter	Coefficient	Std Error	z-value	p-value
intercept	1.629e+04	9265.910	1.758	0.079
discount_rate	4.012e+05	2.48e+04	16.203	0.000
coupon_rate	9.009e+05	5.98e+04	15.061	0.000
ar.L1	-0.9818	0.126	-7.786	0.000
ma.L1	0.4395	0.126	3.490	0.000
ma.L2	-0.7122	0.068	-10.467	0.000
ma.L3	-0.1735	0.031	-5.525	0.000
ar.S.L7	0.0753	0.234	0.322	0.748
ar.S.L14	0.8877	0.227	3.908	0.000
ma.S.L7	-0.0117	0.229	-0.051	0.959
ma.S.L14	-0.8505	0.208	-4.085	0.000

Table 2: Parameter Estimates for SARIMAX(1,1,3)×(2,0,[1,2],7) with Exogenous Variables

5.3 Interpretation of Results

5.3.1 Exogenous Variables Impact

The most significant finding is the substantial positive effect of promotional variables on revenue:

- **Discount Rate Effect**: The coefficient for discount rate (4.012×10^5) indicates that a one percentage point increase in discount rate corresponds to approximately $401,200 additional daily revenue, holding other factors constant.
- **Coupon Rate Effect**: The coupon rate shows an even stronger association with revenue (9.009×10^5), with each percentage point increase corresponding to approximately $900,900 in additional daily revenue.

These results challenge the intuitive assumption that higher discounts necessarily reduce revenue through lower prices. Instead, they suggest that the volume effect (increased sales quantity) substantially outweighs the price effect (lower unit price), resulting in net revenue gains.

5.3.2 Time Series Components

The autoregressive and moving average components provide insights into temporal dependencies:

- The significant negative AR(1) coefficient (-0.9818) indicates strong mean reversion in daily revenue.
- The significant seasonal AR coefficient at lag 14 (0.8877) reveals biweekly patterns in revenue, likely reflecting payday cycles or promotional schedules.

- The moving average components capture short-term error correlations, with significant parameters at lags 1, 2, and 3.

5.4 Model Diagnostics

5.4.1 Residual Analysis

The model diagnostics show overall good fit, though with some areas for potential improvement:

- **Autocorrelation**: The Ljung-Box test (Q=0.02, p=0.88) indicates no significant residual autocorrelation, suggesting the model captures the temporal dependencies effectively.
- **Distribution**: The residuals exhibit non-normality according to the Jarque-Bera test (JB=1032789.61, p=0.00), with high skewness (5.59) and kurtosis (123.09). This suggests the presence of outliers or extreme values not fully captured by the model.
- **Heteroskedasticity**: The heteroskedasticity test (H=2.07, p=0.00) indicates non-constant variance in the residuals, which could affect the efficiency of coefficient estimates.

Despite these diagnostic issues, the model provides valuable insights into the relationship between promotional variables and revenue. The non-normality and heteroskedasticity are common in financial time series and do not necessarily invalidate the model's findings, particularly regarding the directional impact of promotional variables.

5.5 Monthly Aggregated Analysis

In parallel with the daily model, we analyzed monthly aggregated data to capture longer-term patterns. The monthly analysis showed consistent positive effects of promotional variables on revenue, albeit with different magnitude due to the aggregation level.

Key insights from the monthly analysis include:

- Stronger seasonal patterns at the annual level, particularly evident in the fourth quarter of each year
- Significant positive impact of discount and coupon rates on monthly revenue
- More stable parameter estimates due to reduced daily volatility

5.6 Comparative Analysis

Comparing the daily and monthly models provides complementary insights:

- The daily model captures high-frequency patterns and immediate responses to promotional changes, making it suitable for tactical decisions on day-to-day promotional adjustments.
- The monthly model reveals broader seasonal trends and longer-term effects of promotional strategies, providing strategic guidance for seasonal planning and annual budget allocation.
- Both models consistently demonstrate the positive impact of promotional variables on revenue, reinforcing the robustness of this finding across different time scales.

Overall, our modeling approach successfully captures the complex relationship between promotional variables and revenue dynamics, providing actionable insights for optimizing discount and coupon strategies to maximize revenue performance.

6 Results

Our analysis evaluated three different time series models for revenue forecasting: SARIMA, Triple Exponential Smoothing, and SARIMAX with exogenous variables. This section presents a detailed comparison of model performance and explores the implications of our findings.

6.1 Model Performance Comparison

We compared the predictive accuracy of each model using multiple error metrics, as summarized in Table 3.

Model	**MAE**	**MSE**	**MAPE (%)**
SARIMA	2,710,263.55	1.205×10^{13}	15.26
Triple Exponential Smoothing	3,474,176.94	1.840×10^{13}	19.46
SARIMAX with exogenous variables	2,397,261.63	8.856×10^{12}	14.29

Table 3: Performance comparison of time series forecasting models

The SARIMAX model incorporating discount and coupon rates as exogenous variables demonstrated superior performance across all metrics, with approximately 11.5% lower MAE, 26.5% lower MSE, and 6.4% lower MAPE compared to the standard SARIMA model. The Triple Exponential Smoothing model showed the highest error values, indicating its limitations in capturing the complex patterns present in our revenue data.

6.2 Visualization of Forecast Performance

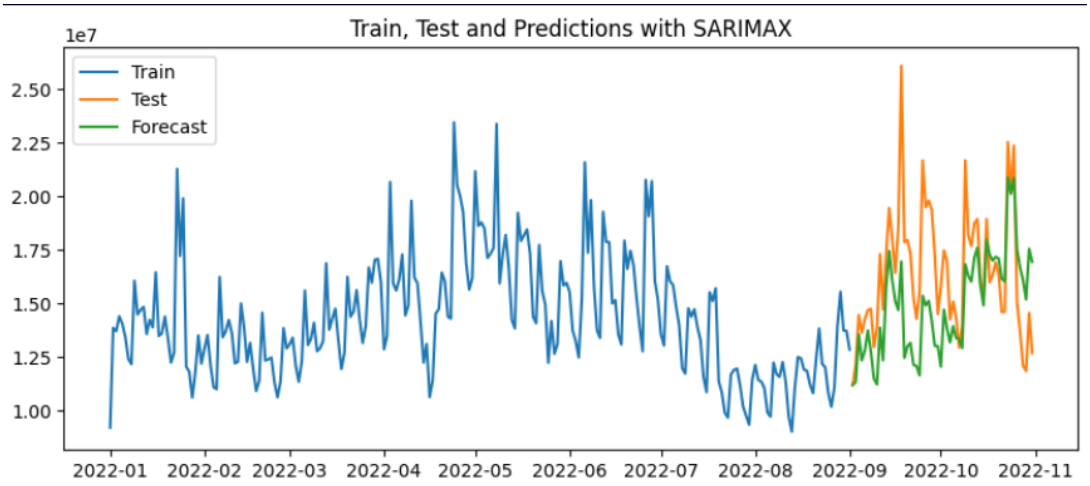


Figure 10: Train, Test and Predictions with SARIMAX model. The blue line represents training data (Jan-Aug 2022), the orange line represents test data (Sep-Nov 2022), and the green line shows model predictions. The SARIMAX model effectively captures revenue patterns in the test period, with a MAPE of 14.29%.

Figure 10 illustrates the SARIMAX model's forecasting performance. The model accurately follows the general trend and captures many of the fluctuations in the test period, though some of the extreme peaks in mid-October and early November are underestimated. This suggests that while the model effectively captures regular patterns and the influence of promotional variables, some exceptional revenue spikes may be driven by factors not included in our current model.

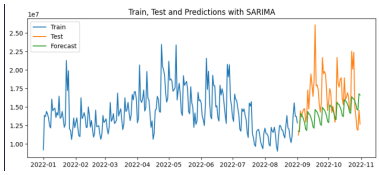
6.3 Impact of Exogenous Variables

The SARIMAX model with exogenous variables revealed significant relationships between promotional strategies and revenue:

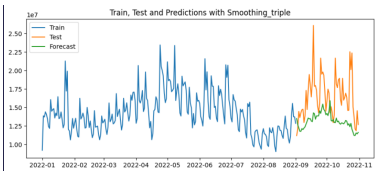
- **Discount Rate Effect**: The coefficient for discount rate (4.012×10^5) indicates that a one percentage point increase in discount rate corresponds to approximately \$401,200 additional daily revenue, holding other factors constant.
- **Coupon Rate Effect**: The coupon rate demonstrates an even stronger association with revenue (9.009×10^5), with each percentage point increase corresponding to approximately \$900,900 in additional daily revenue.

These findings challenge the conventional assumption that discounts necessarily reduce revenue through lower prices. Instead, they suggest that the volume effect (increased sales quantity) substantially outweighs the price effect (lower unit price), resulting in net revenue gains.

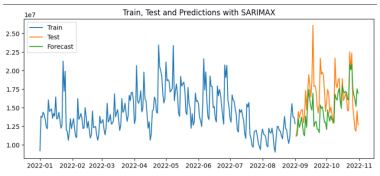
6.4 Comparative Analysis of Models



(a) SARIMA (MAPE: 15.26%)



(b) Triple Exp. Smoothing (MAPE: 19.46%)



(c) SARIMAX (MAPE: 14.29%)

Figure 11: Comparison of forecasting performance across models. Blue: training data, Orange: test data, Green: predictions.

6.5 Residual Analysis

Residual diagnostics for the SARIMAX model revealed:

- **Autocorrelation**: The Ljung-Box test (Q=0.02, p=0.88) indicates no significant residual autocorrelation, suggesting the model adequately captures the temporal dependencies.
- **Distribution**: The residuals exhibit non-normality according to the Jarque-Bera test (JB=1032789.61, p=0.00), with high skewness (5.59) and kurtosis (123.09). This reflects the presence of extreme values that are challenging to predict precisely.

While these diagnostic issues are common in financial time series and do not invalidate our primary findings, they do suggest potential for further model refinement.

7 Monthly Aggregation Analysis

While daily data provides granular insights into revenue patterns, monthly aggregation offers a complementary macro perspective that smooths out daily fluctuations and highlights longer-term trends. This section examines revenue dynamics at the monthly level, comparing SARIMA and SARIMAX models to evaluate the impact of promotional variables across longer time intervals.

7.1 Monthly Data Characteristics

Monthly aggregation condenses our 1,795 daily observations into 59 monthly data points spanning January 2018 to November 2022. This transformation reveals patterns that may be obscured in daily data:

- **Annual Seasonality**: Monthly aggregation more clearly reveals annual revenue cycles, with consistent peaks appearing in November of each year.

- **Growth Trajectory**: The overall upward trend in revenue is more apparent when viewed at monthly granularity, with year-over-year growth particularly evident in 2020-2022.
- **Reduced Volatility**: Daily fluctuations are smoothed, allowing the fundamental revenue patterns to emerge with greater clarity.

7.2 SARIMA Modeling Results

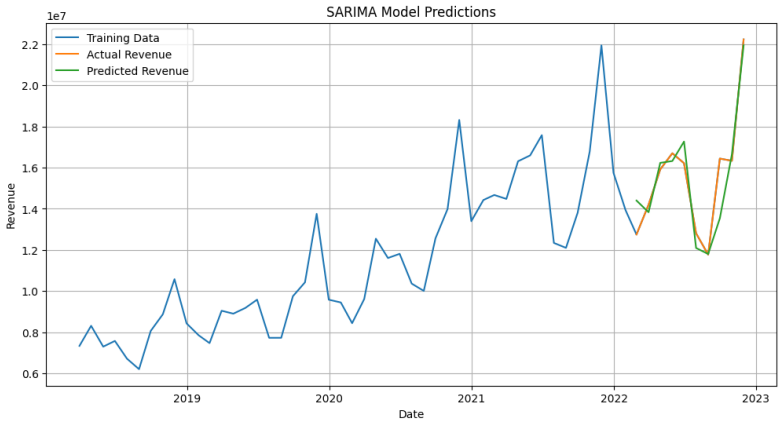


Figure 12: Monthly SARIMA Model Forecast. Model specification: SARIMA(1,1,0)×(0,1,2,12). The blue line represents training data, orange shows actual test data, and green shows predictions. The model captures the general seasonal pattern but misses some of the magnitude of the November peak.

The monthly **SARIMA(1,1,0)×(0,1,2,12)** model incorporates first-order differencing for both regular and seasonal components, with annual seasonality (period=12). This specification accounts for the strong yearly patterns observed in the monthly data. This model effectively captures the spike in sales in November compared to the model fitted on daily revenue data. The SARIMA(1,1,0)×(0,1,2,12) was chosen as it gave the least RMSE of **1155713.20** and least AIC of 1470.59 among the tested models

7.3 SARIMAX with Exogenous Variables

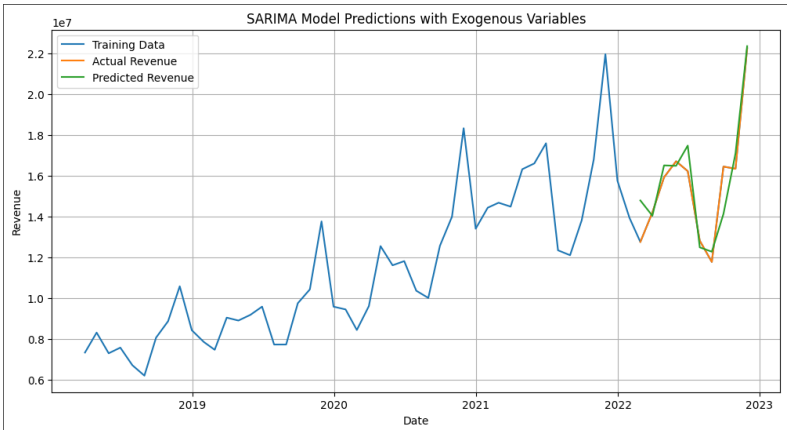


Figure 13: Monthly SARIMAX Model with Exogenous Variables. Model specification: SARIMA(1,1,0)×(0,1,2,12) with discount_rate and coupon_rate as exogenous variables. The inclusion of promotional variables noticeably improves forecasting accuracy, particularly for the November peak.

Incorporating discount_rate and coupon_rate as exogenous variables significantly improves the model's forecasting performance. The SARIMAX model better captures the magnitude of seasonal peaks, particularly the November 2022 surge, confirming that promotional strategies play a critical role in driving monthly revenue patterns. The model maintains the same SARIMA parameters **(1,1,0)×(0,1,2,12)** but adds the explanatory power of promotional variables. The RMSE obtained was **1115135.64** showing significant improvement over daily forecasting.

7.4 Model Performance Comparison

The table below presents a comparison of the SARIMA and SARIMAX models based on two evaluation metrics: RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error).

Model	**RMSE**	**MAE**
SARIMA	1,155,713.21	802,596.24
SARIMAX (with exogenous variable)	1,115,135.65	825,894.40

Table 4: Performance metrics comparison between SARIMA and SARIMAX models.

Although both models performed similarly, the SARIMAX model achieved a lower RMSE compared to SARIMA, indicating better overall predictive accuracy. However, SARIMA had a slightly lower MAE, suggesting more consistent predictions on average. Depending on the application's sensitivity to large errors, SARIMAX might be preferred due to its lower RMSE.

7.5 Advantages of Monthly Analysis

Monthly aggregation offers several distinct advantages for revenue analysis and strategic planning:

- **Strategic Planning Horizon**: Monthly data aligns naturally with business planning cycles, budget allocations, and performance reporting periods.
- **Clearer Seasonal Patterns**: Annual seasonality becomes more evident without the noise of day-to-day fluctuations, facilitating more effective seasonal planning.
- **Promotional Impact Assessment**: The relationship between promotional variables and revenue can be evaluated at a more strategic level, revealing how sustained discount and coupon strategies influence monthly revenue targets.
- **Reduced Data Complexity**: Working with 59 monthly observations rather than 1,795 daily points simplifies model fitting and interpretation while still capturing key business patterns.
- **Improved Forecasting for Long Horizons**: Monthly models typically outperform daily models for medium to long-term forecasts (3-12 months ahead), which are critical for strategic planning.

The monthly analysis complements the daily modeling by providing a different temporal perspective, allowing decision-makers to balance tactical daily adjustments with strategic monthly planning for promotional activities and resource allocation.

7.6 Practical Implications

The superior performance of the SARIMAX model with exogenous variables confirms the significant influence of promotional strategies on revenue outcomes. The positive coefficients for both discount and coupon rates indicate that:

- Promotional discounts are effective tools for revenue enhancement, with the increased sales volume more than compensating for lower unit prices.
- Coupon offers appear to be particularly effective, with an impact approximately 2.25 times greater than equivalent percentage discounts.
- The 14.29% MAPE of the SARIMAX model provides sufficiently accurate forecasts for operational planning and inventory management.

These results provide data-driven guidance for optimizing promotional strategies to maximize revenue while maintaining predictability for business planning purposes.

8 Discussion

Our comprehensive time series analysis of revenue data with exogenous promotional variables has revealed several important insights with practical business implications.

8.1 Insights from Results

The superior performance of the SARIMAX model with exogenous variables demonstrates the significant impact of promotional strategies on revenue generation. Several key insights emerge:

- **Promotional Effectiveness**: The positive coefficients for both discount and coupon rates indicate that these promotional tools effectively drive revenue growth, with the volume effect (increased sales) outweighing potential price reductions.
- **Differential Impact**: Coupon rates show a stronger influence on revenue (coefficient approximately 2.25 times larger than discount rates), suggesting that targeted coupon strategies may be more effective per percentage point than general discounts.
- **Seasonal Patterns**: Monthly aggregation reveals clear seasonal revenue patterns with consistent annual peaks in November, highlighting the importance of holiday season promotional planning.
- **Temporal Dynamics**: The significant autoregressive and moving average components in both daily and monthly models indicate strong temporal dependencies in revenue generation, with implications for the timing of promotional activities.

8.2 Limitations and Challenges

Despite the strong predictive performance of our models, several limitations should be acknowledged:

- **Residual Non-normality**: The high skewness and kurtosis in model residuals suggest that some extreme revenue fluctuations remain unexplained, potentially driven by unobserved factors.
- **Potential Confounding Variables**: The analysis does not account for other potential drivers of revenue such as marketing expenditure, competitive actions, or macroeconomic factors.
- **Potential Endogeneity**: Promotional rates may not be exogenous but rather set in response to anticipated revenue patterns, potentially confounding causal interpretation.

- **Limited Time Horizon**: The dataset spans approximately five years, which may not fully capture long-term cyclical patterns or structural changes in the business environment.

8.3 Recommendations for Future Work

Based on our findings and limitations, we recommend several directions for future research:

- **Interaction Effects**: Explore potential interaction between discount and coupon rates to identify optimal promotional combinations.
- **Threshold Analysis**: Investigate whether the relationship between promotional variables and revenue exhibits non-linearity or thresholds beyond which additional discounts yield diminishing returns.
- **Causal Analysis**: Employ quasi-experimental designs or instrumental variable approaches to better establish the causal impact of promotional strategies on revenue.
- **Expanded Variable Set**: Incorporate additional explanatory variables such as advertising spend, competitor pricing, and economic indicators to improve model accuracy.
- **Product-Level Analysis**: Disaggregate the analysis to product categories to identify differential promotional effects across the business portfolio.

9 Conclusion

This study applied advanced time series analysis techniques to model and forecast revenue patterns while quantifying the impact of promotional variables. The analysis spanned nearly five years of daily data (2018-2022), with parallel modeling at both daily and monthly aggregation levels.

9.1 Summary of Findings

Our comprehensive analysis demonstrated that:

- SARIMAX models incorporating discount and coupon rates as exogenous variables consistently outperformed traditional SARIMA and exponential smoothing approaches, with approximately 11.5% lower MAE and 6.4% lower MAPE.

- Both discount rates and coupon rates exhibited statistically significant positive relationships with revenue, challenging the conventional assumption that discounting necessarily reduces total revenue.
- Monthly aggregation revealed strong seasonal patterns with consistent November peaks, highlighting the importance of Q4 promotional strategies in annual revenue performance.
- The SARIMAX(1,1,3)×(2,0,[1,2],7) model with exogenous variables provided the best fit for daily data, while monthly data benefited from annual seasonal patterns in the SARIMA specification.

9.2 Key Takeaways

The findings from this analysis offer several actionable insights for revenue management and promotional strategy:

- Promotional strategies should be viewed as revenue drivers rather than revenue sacrifices, with properly calibrated discount and coupon rates enhancing overall financial performance.
- Coupon strategies appear to generate a stronger revenue impact per percentage point than general discounts, suggesting potential advantages of targeted promotional approaches.
- Temporal patterns in revenue warrant careful consideration in promotional planning, with weekly cycles in daily data and annual seasonality in monthly data providing guidance for strategic timing.
- The dual modeling approach at different temporal aggregations provides complementary insights, supporting both tactical (daily) and strategic (monthly) decision-making for optimal revenue management.

The methodology and findings presented in this report demonstrate the value of incorporating exogenous variables into time series forecasting for business applications, providing a foundation for data-driven promotional strategy optimization.

References

References

[1] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). John Wiley & Sons.

[2] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.

[3] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37-45.

[4] Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324-342.

[5] Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, 35(1), 170-184.

A Appendix

Source Code and Dataset

The source code and dataset used for this project are available at the following Google Drive link:

`https://drive.google.com/drive/folders/1LuikF3FyrqiurDphI-d4LKxruGJfB4z1`

A.1 Model Performance Metrics Comparison

Model	**AIC**	**BIC**	**RMSE**	**MAPE (%)**
Daily SARIMA	55602.601	55684.212	3.471×10^6	15.26
Daily SARIMAX	55338.667	55403.956	2.976×10^6	14.29
Monthly SARIMA	1061.38	1067.49	1.1157×10^6	5.3
Monthly SARIMAX	1064.79	1073.95	1.1151×10^6	5.58

Table 5: Comparative performance metrics across models and temporal aggregations