Sales Trend Analysis and Forecasting Report

Ting-His Lee 07/30/2025

I PROJECT OVERVIEW

i. Project Goals:

- Analyze past 3 years of sales data
- Identify trends and seasonal patterns
- Build a forecasting model for the next 3 months
- Provide actionable business insights for decision-making

ii. Dataset Description:

The sales dataset used in this project is adapted from the Retail Store Inventory Forecasting Dataset originally published on Kaggle. While the original dataset was designed primarily for inventory forecasting and spans only two years, it provided a valuable foundation with features such as sales quantity, product category, pricing, location, weather conditions, and promotional indicators.

To better align the dataset with the objective of sales trend and forecasting analysis over a three-year period, the following adjustments were made:

- Irrelevant columns (e.g., fields not related to sales performance) were removed to streamline the dataset.
- A third year of data (2024) was synthetically generated using ChatGPT, with careful attention to preserving realistic business patterns.

This extended dataset ensures the presence of meaningful sales patterns for time-series analysis and forecasting, making it suitable for both exploratory insights and predictive modeling.

iii. Meta Data

Column Name	Description
Date	The date of the sales record (daily granularity)
Store ID	Unique identifier for each retail store
Product ID	Unique identifier for each product
Category	Product category (e.g., Electronics, Groceries, Toys)
Region	Geographical region of the store (e.g., North, South)
Units Sold	Number of units sold on that date
Price	Unit price of the product
Discount	Percentage discount applied to the product
Weather Condition	Weather status on that day (e.g., Sunny, Rainy, Cloudy)

Holiday/Promotion	Binary indicator (1 = holiday or promotion, 0 = otherwise)
Competitor Pricing	Average price of the same product from competitors
Seasonality	The seasonal label for the date (e.g., Summer, Winter)

II DATA CLEANING & PREPROCESSING

- **Data Range:** January 2022 – December 2024

- **Feature Engineering**:

  - **Sales** = Price × Units Sold × (1 - Discount%)

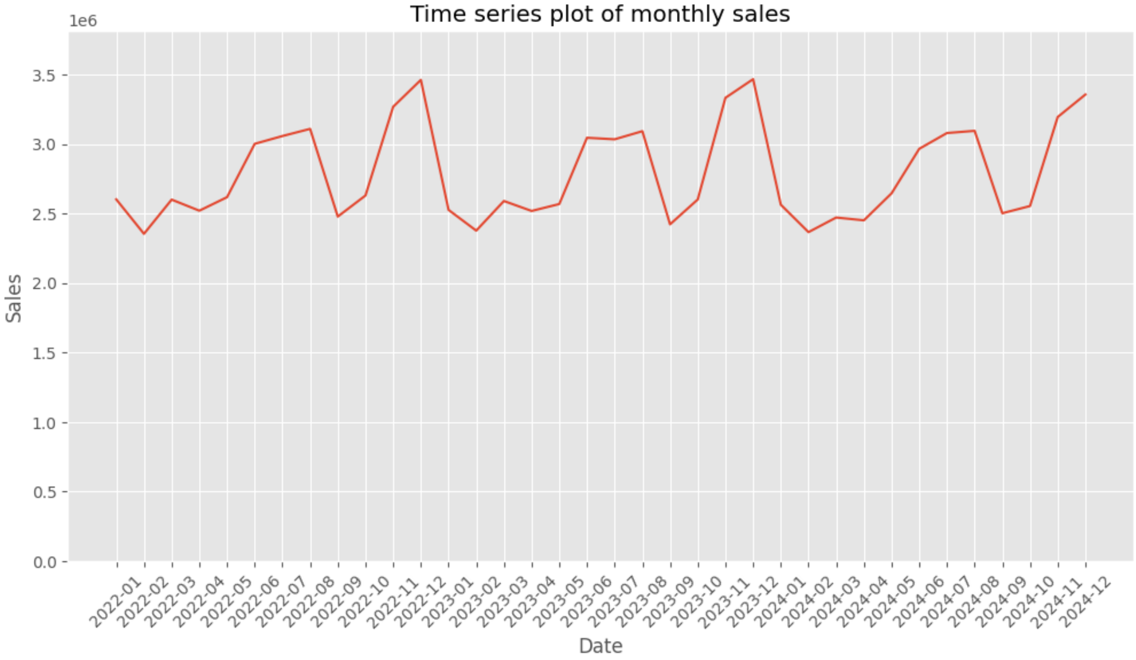
  - **One-Hot Encoding (Dummies):** To explore relationships between categorical variables and sales performance, one-hot encoding was applied to the following categorical columns:

Store ID	Product ID	Category	Region	Weather Condition	Seasonality

- **Monthly Sales Aggregation:** To analyze seasonality and long-term trends, total monthly sales were aggregated by summing daily sales.

III EXPLORATORY DATA ANALYSIS (EDA)

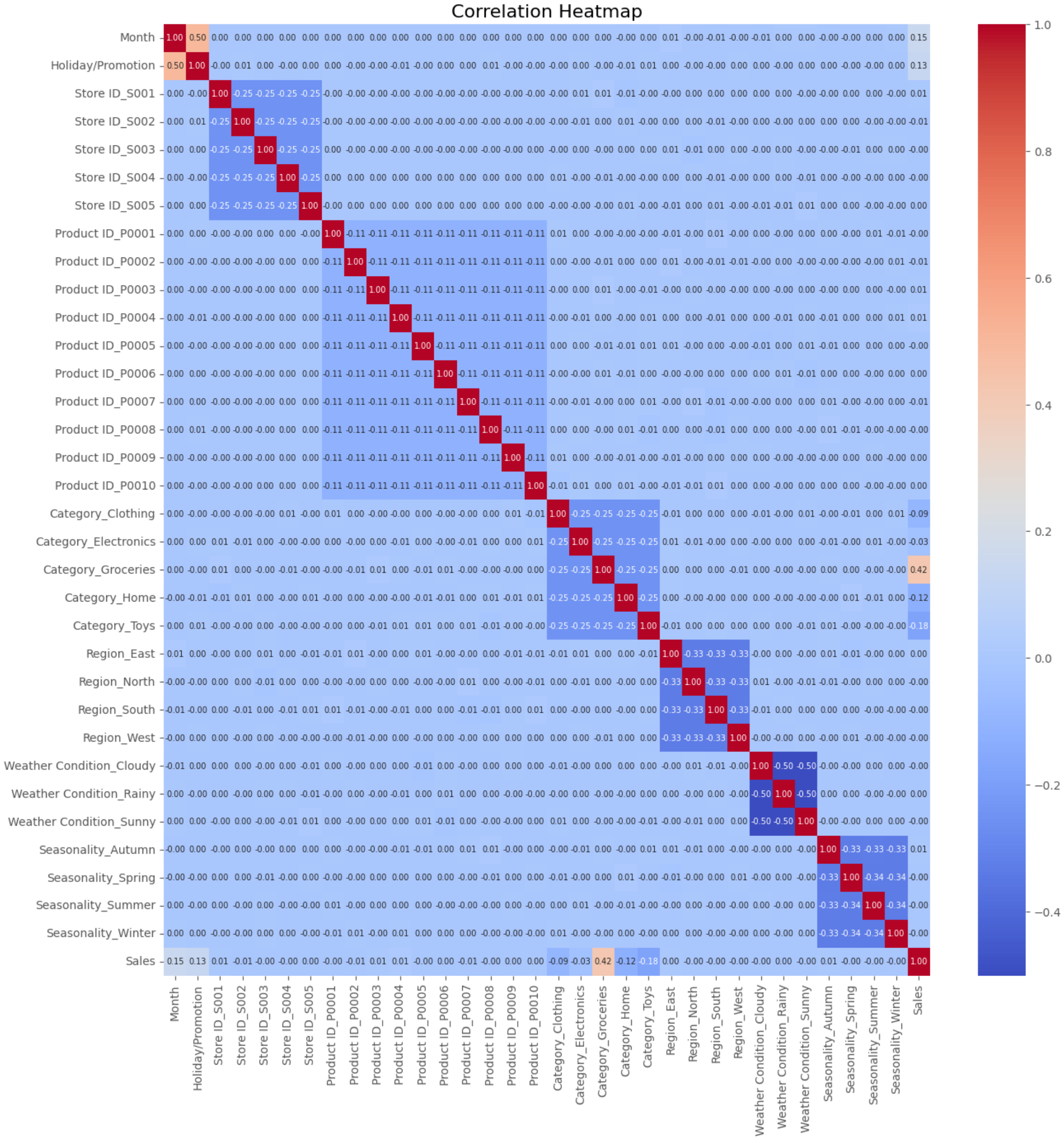
- **Time Series Analysis – Monthly Sales Trend**



This line chart illustrates the monthly sales performance from January 2022 to December 2024. Several key patterns are observable:

- **Seasonality:** Sales exhibit a recurring annual spike in November and December, likely driven by holiday promotions and year-end shopping trends.
- **Summer Boost**: Sales also rise from June to August, likely due to back-to-school promotions**.**
- **Post-Holiday Slowdown:** Sales tend to dip slightly in Q1 each year, following the holiday season.

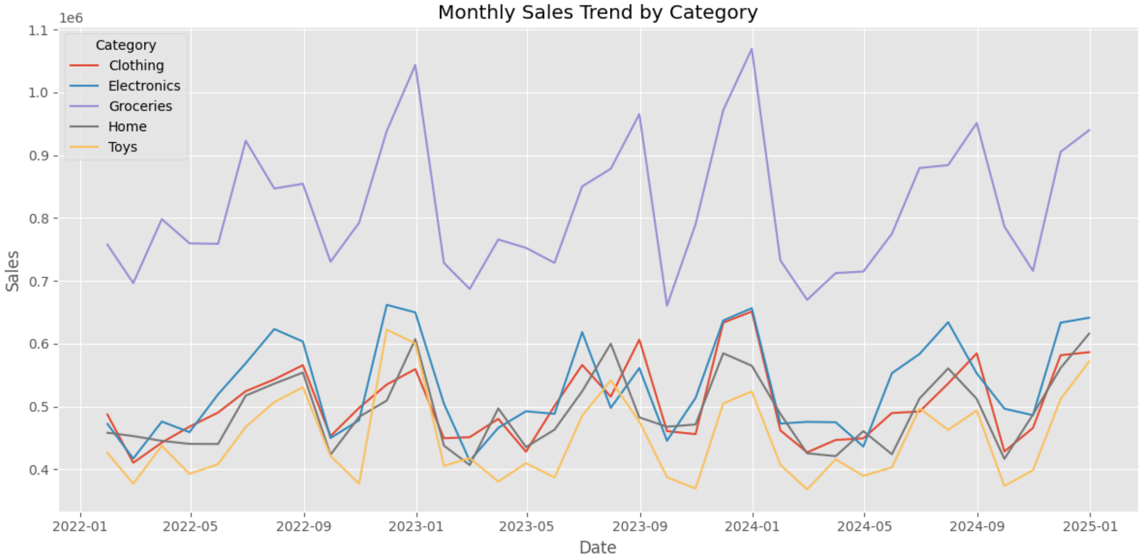
- **Correlation Heatmap Analysis**



Based on the correlation heatmap, the following features show meaningful influence on sales performance:

- **Category**: Grocery has the strongest positive impact, suggesting it's the top revenue-generating product type.
- **Time-related Features**: Month shows that sales vary across the year, indicating seasonal effects.
- **Promotions**: **Holiday/Promotion** periods correlate positively with sales, confirming that marketing campaigns and holiday seasons boost consumer spending.

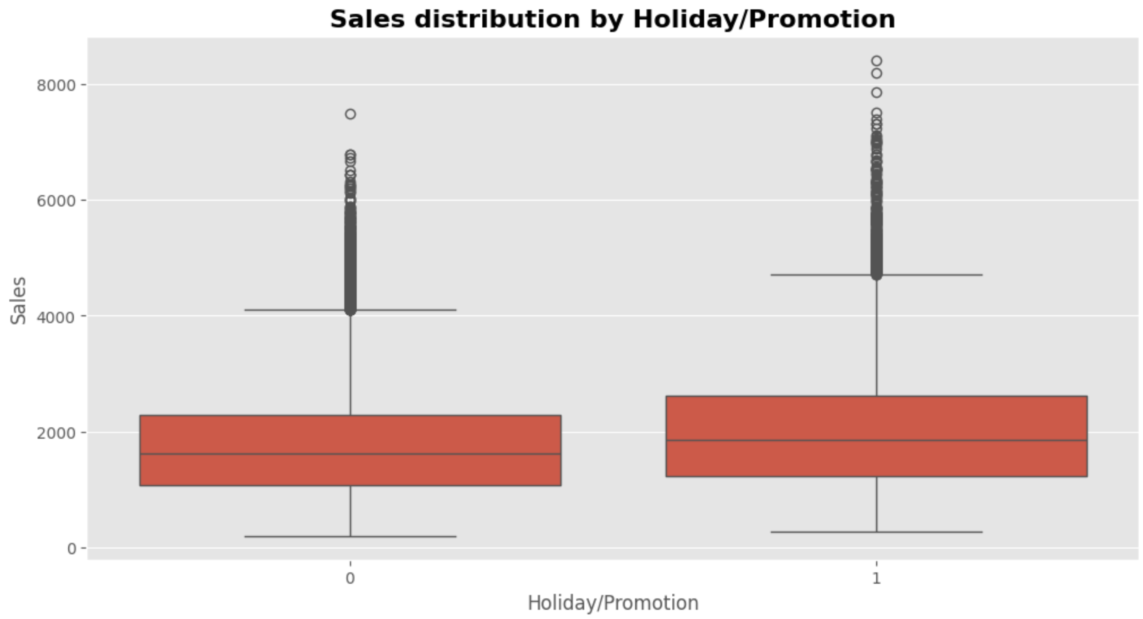
- **Monthly Sales Trend by Category**



The line chart reveals distinct **seasonal trends and performance differences** across product categories:

- All categories exhibit a similar seasonal trend, with noticeable peaks in December and smaller rises during mid-year months (June–August).
- **Groceries** continue to lead in total sales and show the most stable and pronounced seasonal spikes, especially in December.

- **Sales Distribution by Holiday/Promotion**



This graph illustrates the distribution of total sales during regular days (0) and holiday/promotion periods (1). It shows that:

- Sales are generally higher during holidays or promotions (label = 1), with both the median and upper range being noticeably greater than non-holiday periods.
- The interquartile range (IQR) is also wider during holidays, suggesting greater variability — likely due to different campaign scales or product categories.
- More high-value outliers are observed during promotional periods, indicating occasional exceptional spikes in sales.

IV FORECASTING MODEL DEVELOPMENT

Model Choice: ARIMA and SARIMA

Train/Test Split (Train: first 33 months, Test: last 3 months)

Why ARIMA?

To forecast sales for the upcoming three months, I selected ARIMA and SARIMA models due to their proven effectiveness in time series forecasting tasks, especially when working with structured and continuous monthly data.

- **ARIMA (Auto Regressive Integrated Moving Average)** was used as a baseline model. It is suitable for capturing overall trends and short-term dependencies in the data, particularly when there is no strong seasonal pattern.

- However, because the dataset exhibited clear **Seasonal Patterns** — such as significant revenue increases during holiday months like November and December, and slight peaks during the summer months — the **SARIMA (Seasonal ARIMA) model** was ultimately chosen for the final forecast.

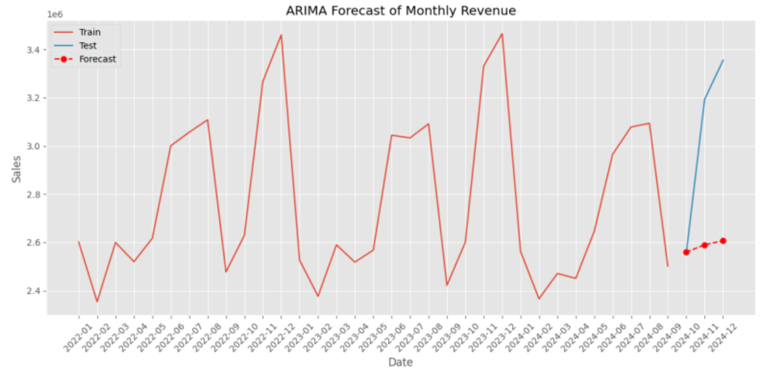
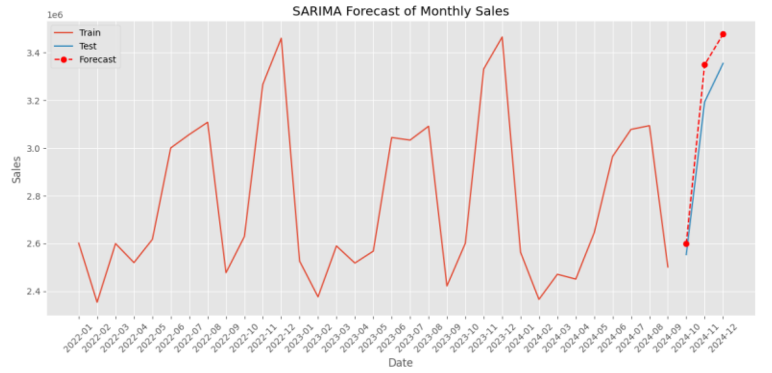
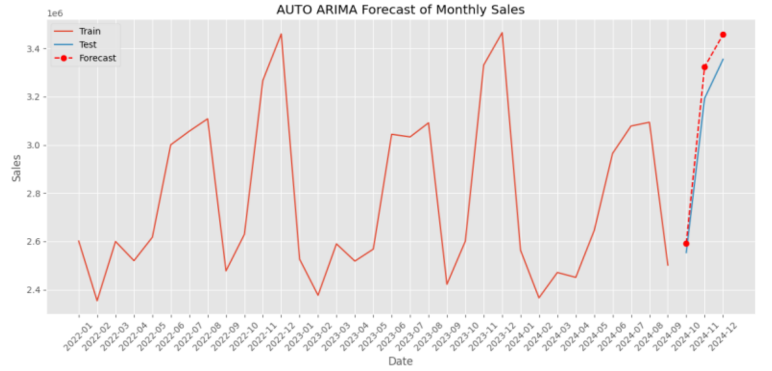
To ensure objectivity and avoid manual tuning bias, I utilized the **auto_arima** function, which uses AIC/BIC criteria to optimize the model's complexity and predictive accuracy without manual intervention, and automatically selects the one that offers the best trade-off between accuracy and simplicity. This allowed for a more efficient and statistically robust model selection process.

V FORECASTING RESULTS AND EVALUATION

i. Time series Modeling

To assess the performance of the forecasting models (ARIMA, SARIMA, and Auto ARIMA), two key evaluation metrics were used:

- MAPE (Mean Absolute Percentage Error):
  - Indicates the average percentage difference between the predicted and actual values.
  - Lower is better — e.g., MAPE of 2.90% means predictions are off by only 2.9% on average.
- RMSE (Root Mean Squared Error):
  - Measures the average magnitude of forecast error in terms of dollars.
  - A lower RMSE means predictions are more accurate.
  - For reference, the average monthly revenue in the dataset is $2.79 million, so an RMSE of $98,620 corresponds to an error of just ~3.5%.

Model		MAPE	RMSE
ARIMA		**13.81%**	**555,069**
SARIMA		**3.45%**	**117,519**
Auto ARIMA		**2.90%**	**98,620**

As visualized in the accompanying figure, the **Auto ARIMA** model outperformed the others across both metrics:

- **ARIMA**, while simple, failed to account for seasonal variations, leading to much higher errors.

- **SARIMA** significantly improved accuracy by modeling repeating seasonal patterns.

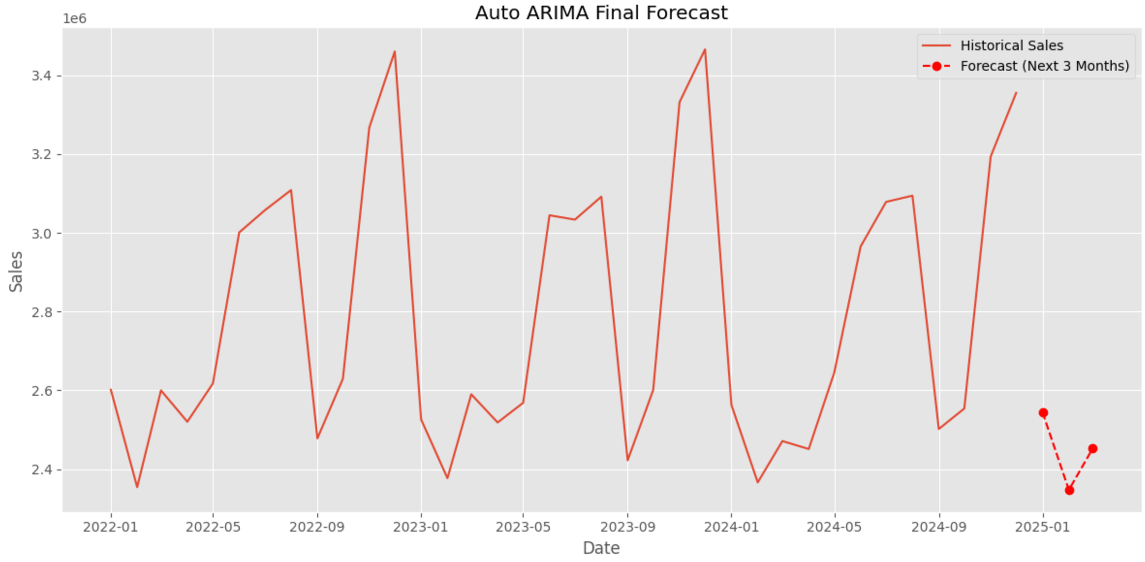
- **Auto ARIMA** automatically selected the best combination of parameters using AIC/BIC, achieving the **most accurate forecast** with the **lowest MAPE and RMSE**.

In conclusion, the Auto ARIMA model achieved the lowest **MAPE of 2.90%** and **RMSE of $98,620**, which is only **3.5% of the average monthly sales** (~$2.79 million). This indicates the model provides both accurate and reliable forecasts for practical use.

ii. Final Forecast for Next 3 Months

Based on the full dataset and the selected Auto ARIMA model, the projected total monthly sales for the next three months are:

Month	Forecasted Sales
Jan 2025	2,544,036
Feb 2025	2,347,014
Mar 2025	2,452,062



The model anticipates a temporary dip in February, consistent with seasonal behavior seen in previous years. This forecast helps the company prepare inventory and promotional strategies for Q1 2025.

VI CONCLUSION

This analysis explored historical sales data through time series modeling and exploratory data analysis to uncover trends and build a reliable forecasting solution. Key insights include:

- **Seasonality plays a crucial role** in sales performance, with consistent peaks during **June–August** and **December**, likely driven by back-to-school, holiday promotions, and end-of-year events.

- **Category-level patterns** show that **Groceries consistently outperform other product groups**, while other categories like Electronics, Clothing, and Home exhibit similar seasonal behavior.

- **Holiday and promotional periods** are associated with significantly **higher sales** and **greater variability.**

- Among the tested models, **Auto ARIMA** outperformed ARIMA and SARIMA, achieving the **lowest error metrics (MAPE 2.90%, RMSE $98,620)**. This level of accuracy—just **~3.5% of average monthly sales**—makes it well-suited for practical forecasting applications.

In summary, using AIC/BIC criteria to optimize the model's complexity and predictive accuracy without manual intervention. The analysis highlights that timing, product category, and promotional strategies are key drivers of sales performance.