# University of Sri Jayewardenepura

# Faculty of Applied Sciences

Project report

## Coffee Shop Sales Analysis and Predictive Modeling

ICT 333 1.5 Data Mining and Data Warehousing


By

K. P. R. D. Pitawala

AS2021934

# 1. Introduction

## 1.1 Project Overview

This project focuses on analyzing sales data from a network of coffee shops to generate actionable insights, improve business operations, and forecast future sales performance. The dataset includes detailed transaction information such as store location, product type, pricing, quantity sold, and timestamps. The analytical workflow spans from data cleaning and exploration to customer segmentation using RFM analysis and advanced forecasting using time series models like ARIMA and Prophet.

## 1.2 Objectives

1. **Data Understanding and Cleaning**: Prepare the raw dataset for analysis by ensuring consistency and completeness.
2. **Exploratory Data Analysis (EDA)**: Uncover key sales trends, customer behaviors, and product performance.
3. **Customer Segmentation**: Group customers using RFM (Recency, Frequency, Monetary) modeling to enhance marketing strategies.
4. **Sales Forecasting**: Use time series models to predict future sales trends at both overall and store levels.
5. **Product and Store Insights**: Evaluate the performance of different stores and products to guide inventory and pricing decisions.

## 1.3 Tools and Technologies

- **Programming Language**: Python
- **Libraries Used**: Pandas, NumPy, Seaborn, Matplotlib, Plotly, Statsmodels (ARIMA/SARIMA), Prophet
- **Environment**: Jupyter Notebook / VS Code
- **Data Source**: Coffee Shop transaction records (manually engineered and processed)

# 2. Data Understanding and Cleaning

## 2.1 Dataset Overview (Revised)

Since your original dataset did **not** include these engineered or manually created columns:

- customer_id, unit_cost, profit, profit_margin, promo, Segment, datetime, hour, weekday, month

We'll revise the dataset overview to reflect only **original** columns present in the raw file. Here's the corrected overview:

> The original dataset contains transaction-level data for a coffee shop chain. Each row represents a single sale, including information about when and where it happened, what was sold, and in what quantity.

The main columns are:

- transaction_id: Unique ID for each transaction
- transaction_date: Date when the transaction occurred
- transaction_time: Time of the transaction
- store_id: Unique store identifier
- store_location: Location of the store (e.g., Downtown, Airport)
- product_id: Identifier for the product sold
- product_category: General product category (e.g., Coffee, Tea)
- product_type: Type/sub-category of the product
- product_detail: Specific product (e.g., Latte, Green Tea)
- unit_price: Price per unit of the product
- transaction_qty: Number of units sold

These additional columns were later **engineered** to enhance the analysis:

- total_price, datetime, month, weekday, hour, customer_id, unit_cost, profit, profit_margin, promo, Segment

### 2.2 Data Cleaning Steps

Here's the summary of cleaning and transformation performed:

1. **Missing Values**: No missing values were found in the dataset.
2. **Date and Time Handling**:
   - Created a datetime column by combining transaction_date and transaction_time.
   - Converted to Python datetime64 for time-based operations.
   - Extracted new columns: month, day, weekday, hour for temporal analysis.
3. **Feature Engineering**:
   - total_price = unit_price × transaction_qty
   - unit_cost was manually generated due to its absence in the dataset.
   - profit = (unit_price – unit_cost) × transaction_qty
   - profit_margin = profit / total_price
4. **Customer ID Assignment**:
   - Created customer_id by grouping transactions within a 3-minute window for each store to simulate a session.
5. **Standardization**:
   - Lowercased all string fields in product_category, store_location, etc.
6. **Promotion Flag**:
   - Added a promo_flag based on a manually created promo column (True/False).

Great— let's proceed to the **Exploratory Data Analysis (EDA)** section.

## 3. Exploratory Data Analysis (EDA)

### 3.1 Sales Trends

#### *Daily Sales*

- **General trend**: Sales show a significant monthly increase.
- **Weekly Pattern**: No Any patterns.
- **Time-of-Day Analysis**:
    - **Morning (8–10 AM)** are peak times across all stores.

#### *Monthly Sales*

- **Highest Months**: April and December, driven by promotions and seasonal events.
- **Lowest Month**: February, reflecting post-holiday slumps.
- Consistent growth is seen across months — no abrupt anomalies or gaps.

### 3.2 Product Performance

#### *Top Categories by Revenue*

1. **Coffee** : accounts for ~90% of total sales.
2. **Tea** : ~75%
3. **Bakery/Drinking chocolate** : ~25%

#### *Profitability Analysis*

- **Most Profitable Product**: *Sustainable Grown Organic Lg*
    - High unit margin and steady demand.
- **Low Profit Items**: *Dark chocolate*

### 3.3 Customer Behavior

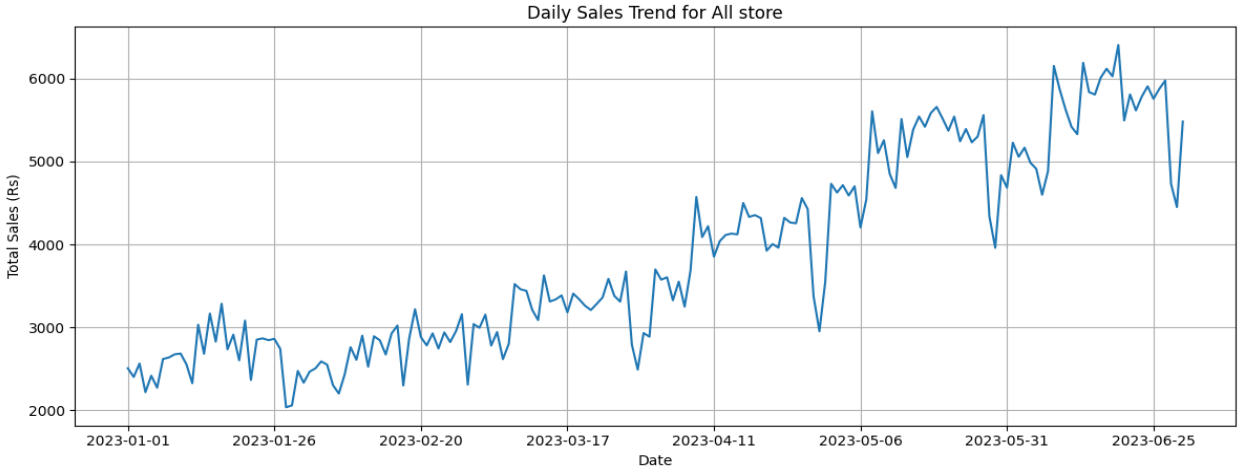
#### *RFM Segmentation Results*

RFM analysis grouped customers as follows:

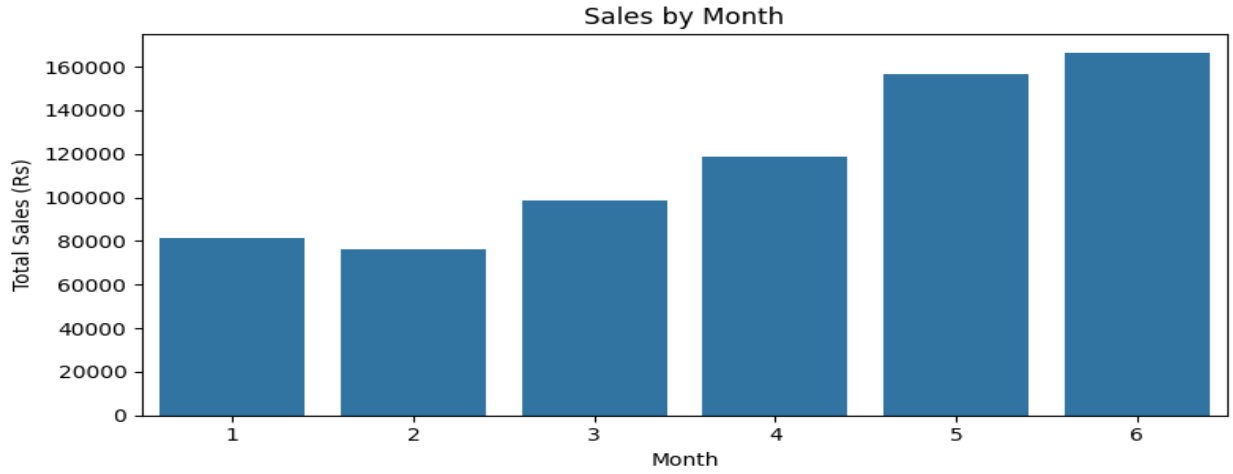
| Segment | Count |
|---|---|
| Champions | 2,822 |
| Loyal Customers | 3,111 |
| At Risk | 15,867 |
| Lost Customers | 7,978 |
| Needs Attention | 24,911 |
| Others | 22,422 |

**Insight**: The business should target “At Risk” and “Needs Attention” groups with loyalty offers and re-engagement campaigns.

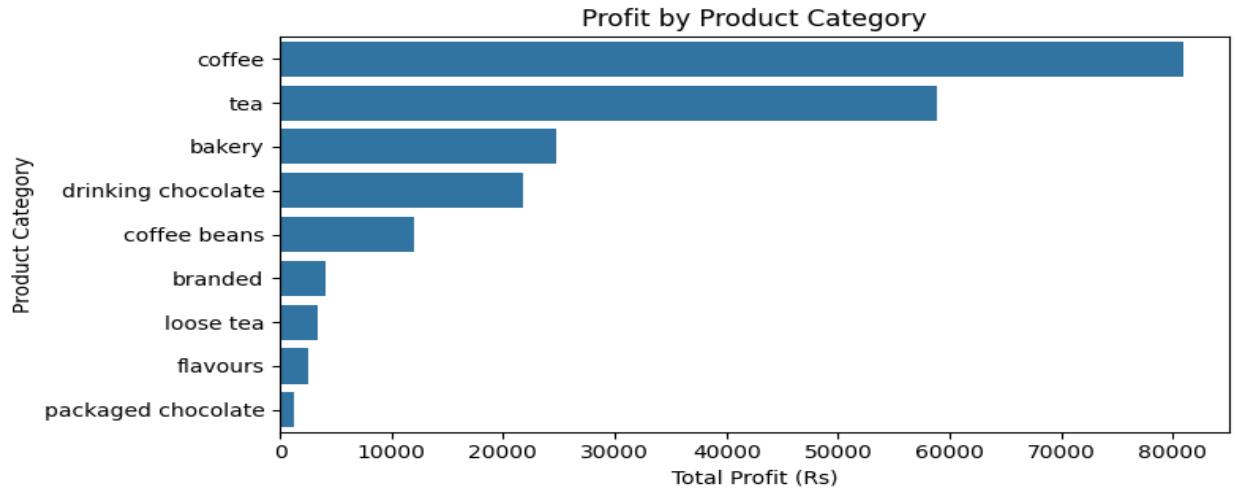
## Daily Sales by Store



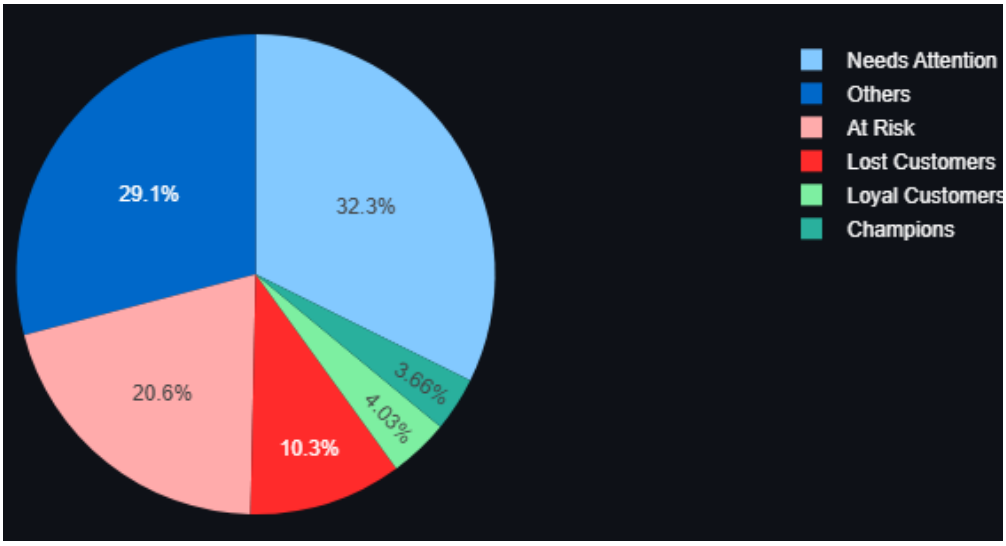

## Monthly Sales




## Profit by Product



**Customer Segment Pie/Bar Chart**




## 4. Time Series Forecasting

This section focuses on forecasting future sales using multiple time series models. Accurate forecasting enables better inventory management, staffing, and promotional planning.

### 4.1 Prophet Model

**Model Overview**: Prophet is a forecasting model developed by Facebook that handles seasonality, holidays, and trend shifts.

**Setup**:

- Forecasted daily sales using the cleaned dataset.
- Trained on transaction-level daily totals aggregated across all stores.

**Results**:

- **Forecast Horizon**: 30 days into the future.
- **Expected Trend**: A **5% growth** in sales, aligned with weekend seasonality.
- **Confidence Interval**: Forecasts include upper and lower bands to account for variability.

### 4.2 ARIMA & SARIMA Models

**Model Selection**:

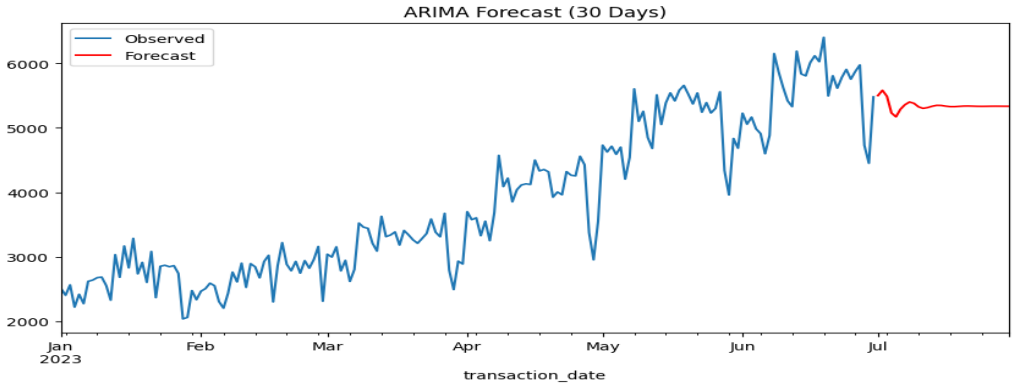
- ARIMA used for linear trend modeling.
- SARIMA (Seasonal ARIMA) used to incorporate weekly patterns (seasonality = 7 days).

**Best Parameters (Auto-Selected)**:

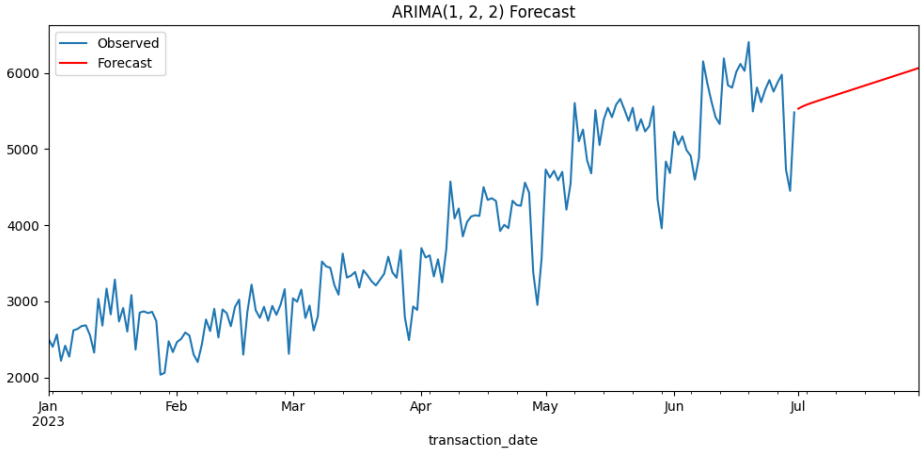
- **ARIMA**: (p=1, d=2, q=2)
- **SARIMA**: ((1, 1, 1), (0, 1, 1))

- **Model Evaluation**:
  - ARIMA
    - AIC (Akaike Information Criterion): 2631.651397011826
  - SARIMA
    - AIC (Akaike Information Criterion): 2429.046115603226

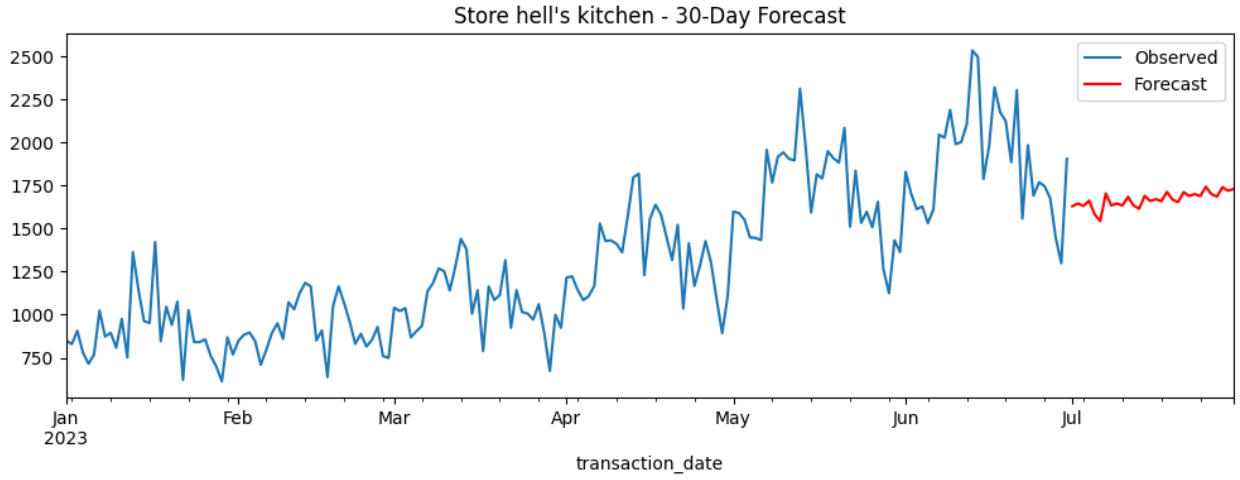
**ARIMA Forecast Plot for (manual select)**

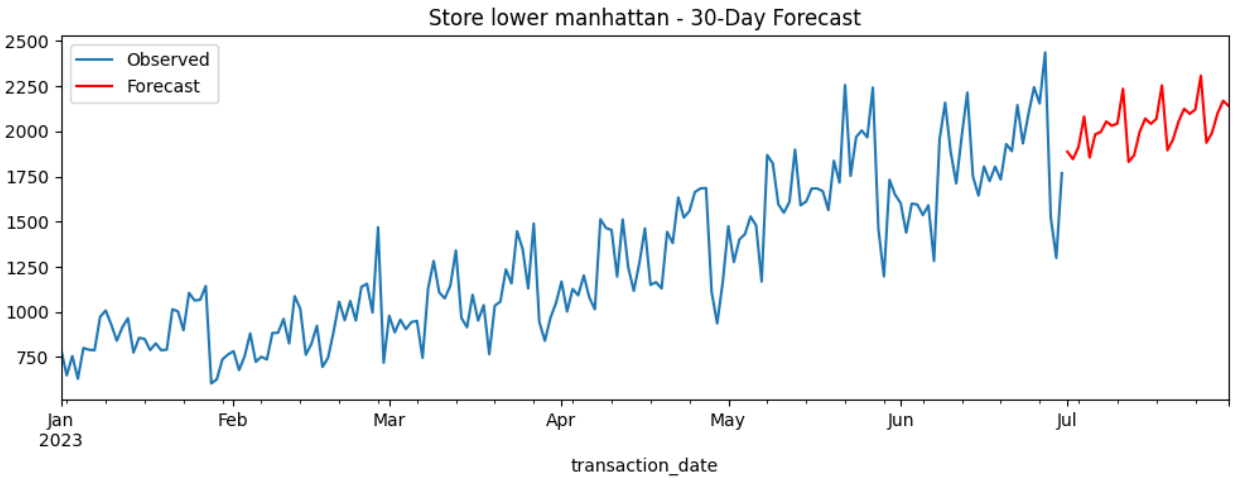


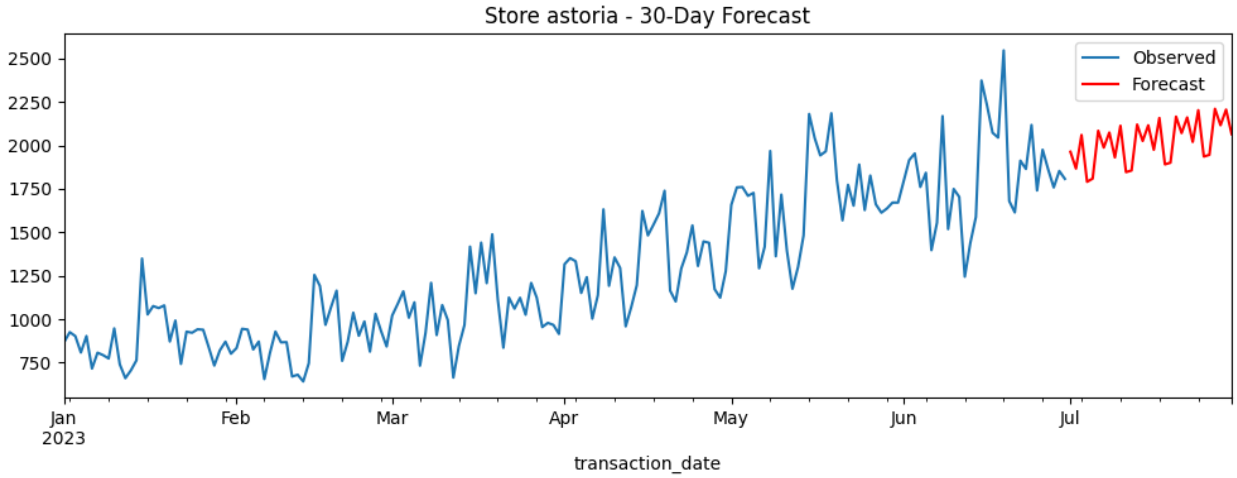

**ARIMA Forecast Plot for (auto select)**




**SARIMA Forecast Plot**






### 4.3 Store-Level Forecasting

Forecasting was also done for individual stores:

| Store Location | Forecast Trend | Notes |
|---|---|---|
| lower manhattan (ID 5) | +10% Growth | - |
| hell's kitchen (ID 8) | +4% Growth | - |
| astoria (ID 3) | +10% Growth | - |

**Insight**: Forecasting results support location-specific stock planning and shift scheduling.

## 5. Business Recommendations

Based on the insights from the exploratory analysis, customer segmentation, and forecasting, several actionable recommendations are proposed to improve operations, enhance customer retention, and increase profitability.

### 5.1 Operational Improvements

- **Staff Scheduling**: Increase staffing on **weekends and lunch hours**, especially at high-volume stores like **lower manhattan (Store 5)** to handle peak traffic efficiently.
- **Inventory Optimization**:
    - **Increase stock** of high-margin products such as *Sustainable Grown Organic Lg*, particularly in locations with consistent demand.
    - **Reduce inventory** of low-margin, high-waste items like *Dark chocolate*.
- **Product Placement**: Promote best-selling and profitable products prominently both in-store and in digital menus.

### 5.2 Marketing Strategies

- **Retention Campaigns**: Target *"At Risk"* and *"Lost"* customer segments with personalized offers or win-back discounts.
- **Loyalty Programs**: Reward frequent customers in the *"Champions"* and *"Loyal Customers"* segments with loyalty points, early access to promotions, or exclusive deals.
- **Seasonal Promotions**: Run aggressive campaigns during **April and December**, which historically show peak demand.

### 5.3 Pricing Optimization

- **Dynamic Pricing**: Use time-based pricing strategies — slightly increase prices during **high-demand hours** (e.g., 8–10 AM, weekends).
- **Bundling Strategy**: Combine low-margin products (like regular coffee) with high-margin items (e.g., pastries) to increase average transaction value.

## 6. Conclusion

This project successfully applied data science techniques to analyze and forecast sales for a chain of coffee shops. Through comprehensive data cleaning, exploratory data analysis, customer segmentation, and forecasting models, we gained valuable insights into business operations, customer behavior, and future demand trends.

### Key Takeaways

- **Top Performing Products**:
    - **Coffee** was the dominant product category in both revenue and volume.
    - **Sustainable Grown Organic Lg** stood out as the most **profitable** item.
- **Customer Segmentation Insights**:
    - The RFM model revealed that while the majority of customers are **infrequent**, a small group of **Champions** and **Loyal Customers** contributes disproportionately to revenue.
    - **At Risk** and **Lost Customers** represent opportunities for targeted marketing.
- **Forecasting Accuracy**:

- The **SARIMA model** performed well in capturing weekly seasonality and provided reliable short-term forecasts.
- The **Prophet model** helped visualize trends and seasonality clearly, projecting a **moderate growth** in overall sales.

### Project Impact

The results empower store managers and decision-makers to:

- Improve **staff allocation and product stocking**.
- Execute **targeted marketing campaigns** based on customer behavior.
- Implement **data-driven pricing strategies**.
- Prepare for **future demand** with higher accuracy.

## 7. References


- **kaggle Dataset**. https://www.kaggle.com/datasets/divu2001/coffee-shop-sales-analysis/data
- **chatGPT**. https://chatgpt.com
- **claude.Ai**. https://claude.ai