Supermarket Sales Trend Analysis

A Final report for the BDM capstone Project

Submitted by

Name: **Pratyush Kala**

Roll number: **23f1002374**

IITM Online BS Degree Program,
Indian Institute of Technology, Madras, Chennai
Tamil Nadu, India, 600036

Contents

1 Executive Summary

This report presents a comprehensive analysis of supermarket sales trends using historical data from three branches in Myanmar, covering January to March 2019. The project focused on addressing critical business challenges such as improving profit margins, optimizing inventory, understanding customer satisfaction, and forecasting demand.

Through detailed data cleaning and preparation, advanced analytics techniques were applied, including Pareto analysis, customer feedback evaluation, trend analysis, and time series forecasting (ARIMA and SARIMA models). The findings reveal that revenue is evenly distributed across products and customers, highlighting the importance of maintaining broad-based marketing and balanced inventory strategies. Customer feedback emphasized areas for improvement, particularly in the Home & Lifestyle category, while trend analysis identified key sales peaks mid-month and seasonal fluctuations across the quarter.

Importantly, forecasting models provided actionable insights into expected demand patterns, supporting data-driven decision-making for inventory, staffing, and promotions. This report offers strategic recommendations to enhance operational efficiency, boost customer satisfaction, and ultimately improve profitability, reflecting the work completed during this final phase and marking a significant advancement beyond the initial project proposal.

2 Proof of Originality

The organization that I am analyzing for this project is a supermarket chain based in Myanmar. The data has been taken from Kaggle.com from a post published by an author called Aung Pyae. The supermarket operates across **three branches** located in different areas of the city.

Following is the link to the dataset that is being analyzed:

https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

3 Meta Data and Descriptive Statistics

Metadata Summary:

The dataset used for this analysis was sourced from Kaggle, titled *Supermarket Sales Data* by Aung Pyae. It contains sales transaction records from three branches of a supermarket chain in Myanmar, covering the period from January to March 2019. The dataset includes 1000 records (rows) and 17 features (columns) describing various aspects of the sales, such as product details, customer information, financial metrics, and customer feedback

Attribute	Details
Dataset Source	Kaggle: Supermarket Sales Data by Aung Pyae
Time Period Covered	January 2019 – March 2019
Geographic Coverage	3 branches in Myanmar (Yangon, Naypyitaw, Mandalay)
Total Records	1000 rows
Total Features	17 columns
Data Types	**Numeric:** (Unit Price, Quantity, Total, COGS, Gross Income, Rating), **Categorical:** (Branch, City, Customer Type, Gender, Product Line, Payment), Date/Time (Date, Time)

Table 1: Dataset information

<u>Attribute Information</u>

- **Invoice ID**: Unique identifier for each computer-generated sales invoice.
- **Branch**: The specific branch of the supermarket where the purchase was made (Branches A, B, and C).
- **City**: The city location of the supermarket.
- **Customer Type**: Classification of customers—'Member' for those with a membership card and 'Normal' for non-members.
- **Gender**: The customer's gender.
- **Product Line**: Category of purchased products (e.g., Electronic Accessories, Fashion Accessories, Food and Beverages, Health and Beauty, Home and Lifestyle, Sports and Travel).
- **Unit Price**: Price per unit of the product in USD.
- **Quantity**: Number of units purchased by the customer.
- **Tax**: A 5% tax applied to the total purchase.
- **Total**: Final purchase amount including tax.
- **Date**: The transaction date (records range from January to March 2019).
- **Time**: Time of purchase (between 10 AM and 9 PM).
- **Payment**: Method of payment used—Cash, Credit Card, or E-wallet.
- **COGS (Cost of Goods Sold)**: The cost incurred by the store for the products sold.
- **Gross Margin Percentage**: The percentage representing gross margin.
- **Gross Income**: Profit earned before expenses, excluding tax.

- **Rating**: Customer's satisfaction score based on their overall shopping experience (scale of 1 to 10).

Descriptive Statistics (Numerical Features):

Quick summary statistics of the dataset:

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Unit Price (USD)	1000	55.67	26.49	10.08	32.88	55.23	77.94	99.96
Quantity	1000	5.51	2.92	1.00	3.00	5.50	8.00	10.00
Tax (5%) (USD)	1000	15.37	11.71	0.51	5.92	12.09	22.45	49.65
Total (USD)	1000	322.97	245.89	10.68	124.42	253.85	471.35	1042.65
COGS (USD)	1000	307.59	234.18	10.17	118.50	241.76	448.91	993.00
Gross Margin (%)	1000	4.76	0.00	4.76	4.76	4.76	4.76	4.76
Gross Income (USD)	1000	15.37	11.71	0.51	5.92	12.09	22.45	49.65
Rating (1–10 scale)	1000	6.97	1.72	4.00	5.50	7.00	8.50	10.00

Table 2: Summary Statistics

Categorical Feature Distribution:

The distribution of key categorical features is as follows:

Branch	Count
A	340
B	332
C	328

Table 3: Branch count

Payment Method	Count
Cash	345
Credit Card	344
E-wallet	311

Table 4: Payment information

Product Line	Count
Food and Beverages	178
Electronic Accessories	174
Fashion Accessories	170
Health and Beauty	166
Home and Lifestyle	160
Sports and Travel	152

Table 5: Product Distribution

Missing Values Summary:

Feature	Missing Values
All Features	0 (No missing values found)

Table 6: Missing Values Summary

Summary

The dataset provides a rich combination of transactional, financial, and customer feedback data. The numerical statistics reveal a broad range of unit prices, quantities, and customer satisfaction ratings, while categorical distributions highlight the supermarket's balanced operations across branches, product lines, and payment methods. This foundation sets the stage for in-depth analysis of trends, profitability, and customer behavior.

4 Detailed Explanation of Analysis Process

Tools used for analysis:

- Python and Python libraries
- Excel
- VS code and google colab

1. **Data Preprocessing and Cleaning**

The dataset did not have any missing values; however, it contained some values that required cleaning, particularly in the Date column. This column contained dates in inconsistent formats:

- Some dates were formatted as mm/dd/yyyy (e.g., 1/27/2019 for January 27, 2019)
- Others used dd-mm-yyyy format (e.g., 1-05-2019, which Excel could misinterpret as May 1, 2019 instead of January 5, 2019)

This inconsistency posed a significant challenge because:

- The dataset only covers January to March 2019, so any dates appearing as April or later would clearly indicate format misinterpretation
- Automated processing would yield incorrect results without proper format standardization

Hence python was used extract the datetime from the column and return a corrected dataset from which the value could be extracted.

Another preprocessing step taken was deriving the Month_Year column that had to be used for future analysis like monthly sales analysis.

- The column was created by transforming the cleaned **Date** values into a standardized mmm-yyyy format (e.g., Jan-2019)
- This transformation enabled:
  - Groupings of sales by month for clearer trend analysis
  - Comparison of sales across time periods
  - Simplification of time-series visualizations and summaries

This was again done using python.

After this some consistency measures were taken, like checking for duplicated records to avoid double counting (done using python), analysis for inconsistent entries and typos (done in excel) and checking for logical discrepancies in data (done via excel). The dataset however was free from all of these issues as well.

2. Pareto Analysis

The goal of Pareto analysis in this project was to identify the key products and customers contributing the most to total revenue and returns. This follows the 80/20 rule, which states that a small percentage of products or customers often generate a large percentage of the revenue.
Pareto analysis was done exclusively using excel.

Methodology:

1. **Data Preparation:**
   - Extracted total sales revenue per product and per customer.
   - Sorted products/customers in descending order based on total revenue.
2. **Pareto Principle (80/20 Rule) Application:**
   - Calculated cumulative revenue contribution.
   - Identified the percentage of top products/customers contributing to 80% of total revenue.
3. **Insights from Pareto Analysis:**
   - The 80/20 principle of the Pareto analysis did not hold true in this analysis. This means the revenue is spread pretty evenly between different customers and products.
   - Neither a small percentage of products nor of customers was responsible for the majority of purchases/revenue.
4. **Visual Representation:**
   - Plotted Pareto chart to show cumulative revenue contribution and determine the cut-off for 80% contribution.
   - Created separate charts for sales contribution.

3. Customer Feedback Analysis

Another part of the analysis process was analyzing customer feedback. Here the customer feedback was analyzed on a scale of 1-10 with respect to the product(s) they bought.

Customer feedback was analyzed in excel.

Methodology:

1. **Data Preparation:**
    1. Extracted mean rating per product line.
    2. Sorted product lines in descending order based on average rating.
2. **Insights from Customer Feedback Analysis:**
    1. The analysis shows how the customer feels about the product purchased and the services provided by the supermarket.
    2. It shows us that the Home and Lifestyle product line requires some improvement.
    3. Branch B has the worst customer response.
3. **Visual Representation:**
    1. Plotted the column chart showing the product lines and their corresponding average rating.
    2. Plotted a clustered column chart showing product lines and their average rating in all 3 branches.
    3. Visual representation makes it easy to compare the different products considered.

4. Trend Analysis

Understanding trends is critical for identifying seasonal patterns, branch performance, and operational efficiencies. This analysis examines monthly sales, daily sales and profit (gross income) across branches to uncover actionable insights for strategic decision-making.

Methodology:

1. **Data Preparation:**
    - Extracted daily and monthly sales totals along with gross income for January–March 2019.
    - Segregated data by branch (A, B, C) to compare performance.
2. **Trend Identification:**
    - Calculated monthly aggregates to assess seasonality.
    - Analysed daily fluctuations in January to pinpoint peak and low-sales periods.
3. **Insights from Trend Analysis:**
    - **Monthly Trends:** January's higher revenue suggests high demand post holidays while February's dip suggests seasonal lulls.
    - **Branch Performance:** Branch C outperformed in high-revenue months (January and March), highlighting potential best practices.
    - **Daily Patterns:** Sales peaked mid-month for all the months.
4. **Visual Representation:**

- **Line chart** of monthly sales with branch breakdowns.
- **Bar chart** of top/low sales days for operational planning.

5. <u>Forecasting</u>

Another part of the analysis focused on **sales forecasting**. Here, historical sales data was used to predict future sales trends using ARIMA modeling, helping anticipate demand patterns and support data-driven inventory and business decisions. ARIMA is well-suited for univariate time series with patterns, while SARIMA captures seasonality.

Methodology:

1. Data Preparation:

- Transformed daily sales data into a time series format with consistent frequency.
- Checked for missing values and ensured data continuity from January to March 2019.

2. Stationarity Testing:

- Applied Augmented Dickey-Fuller (ADF) test to determine stationarity.
- Differenced the series once (d=1) to make it suitable for ARIMA modelling.

3. Model Identification:

- Analysed Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots to estimate AR (p) and MA (q) parameters.
- Evaluated ARIMA models (e.g., ARIMA(1,1,1), ARIMA(0,1,1)) to predict forecast.
- Evaluated SARIMA models to predict forecast.

4. Forecasting and Interpretation:

- Forecasted daily sales for the next 30 days.
- Identified expected peak and low sales periods to assist in inventory and staffing decisions.
- Given the short time period, forecasts should be interpreted cautiously; more data over longer periods would improve accuracy.

5. Visual Representation:

- Line chart comparing actual vs forecasted sales.
- Shaded confidence intervals to illustrate prediction uncertainty.

5 Results and Findings

<u>Pareto Analysis</u>:

- The 80/20 principle of the Pareto analysis did not hold true in this analysis. This means the revenue is spread pretty evenly between different customers and products.
- Neither a small percentage of products nor of customers was responsible for the majority of purchases/revenue.

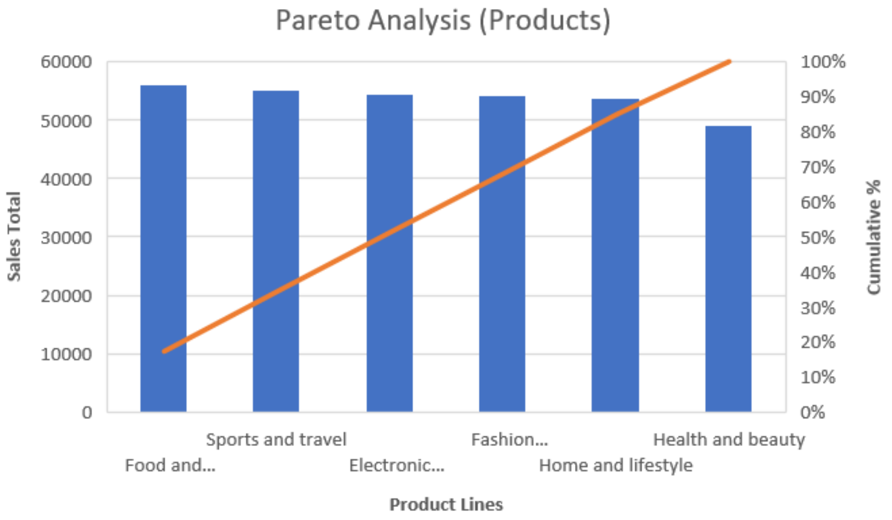


fig 1. Pareto Analysis (Products)

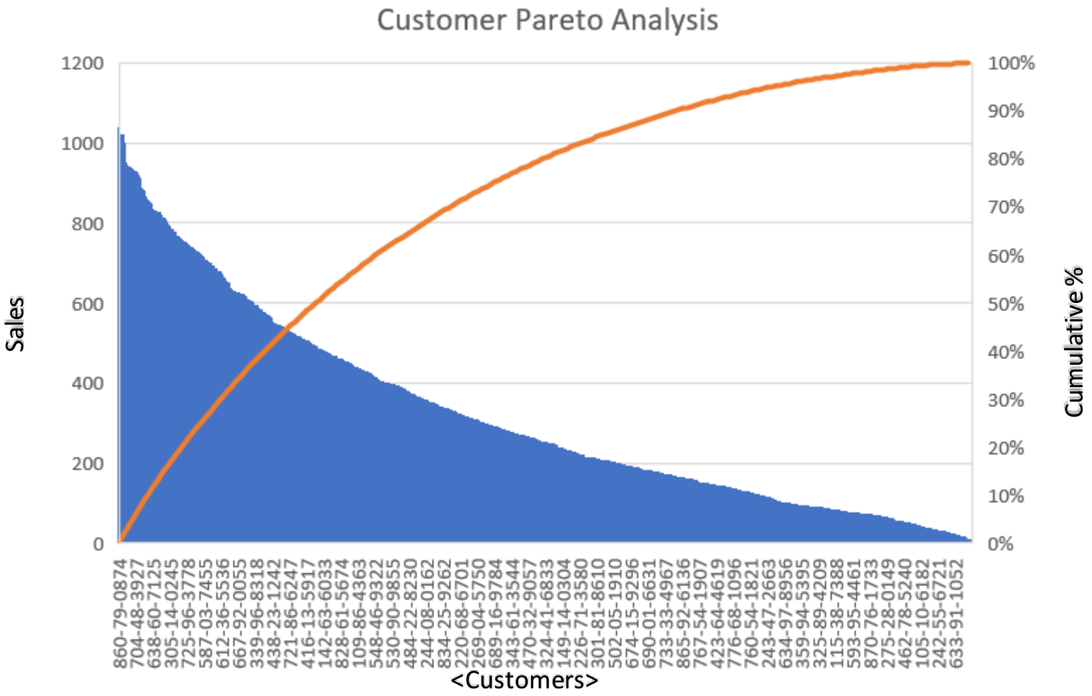


fig 2. Pareto Analysis (Customers)

Interpretation:

Revenue is spread evenly among the 6 different product lines and also fairly evenly among different customers:

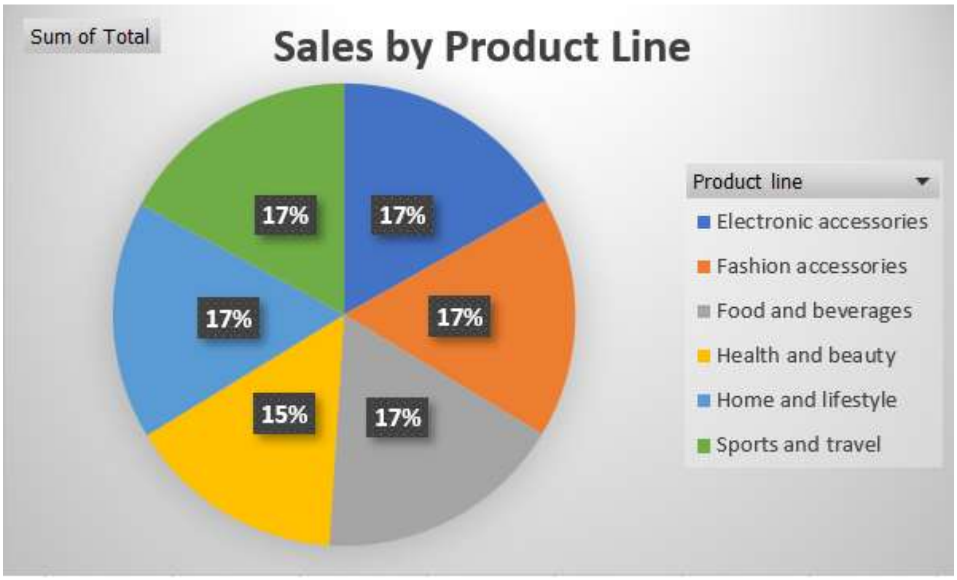


fig 3. Sales by Product line

Feedback Analysis:

The feedback analysis was done using a rating given by the customers on a scale of 1 to 10 and they were compared in consideration to the product line of the customer's purchase:

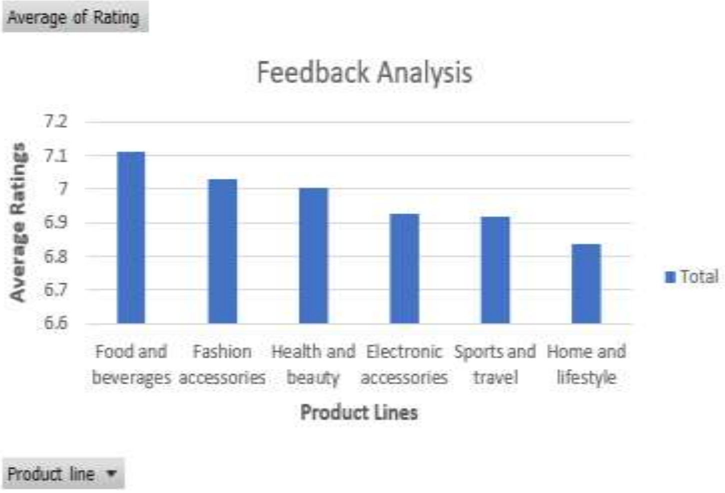


fig 4. Customer Feedback Analysis

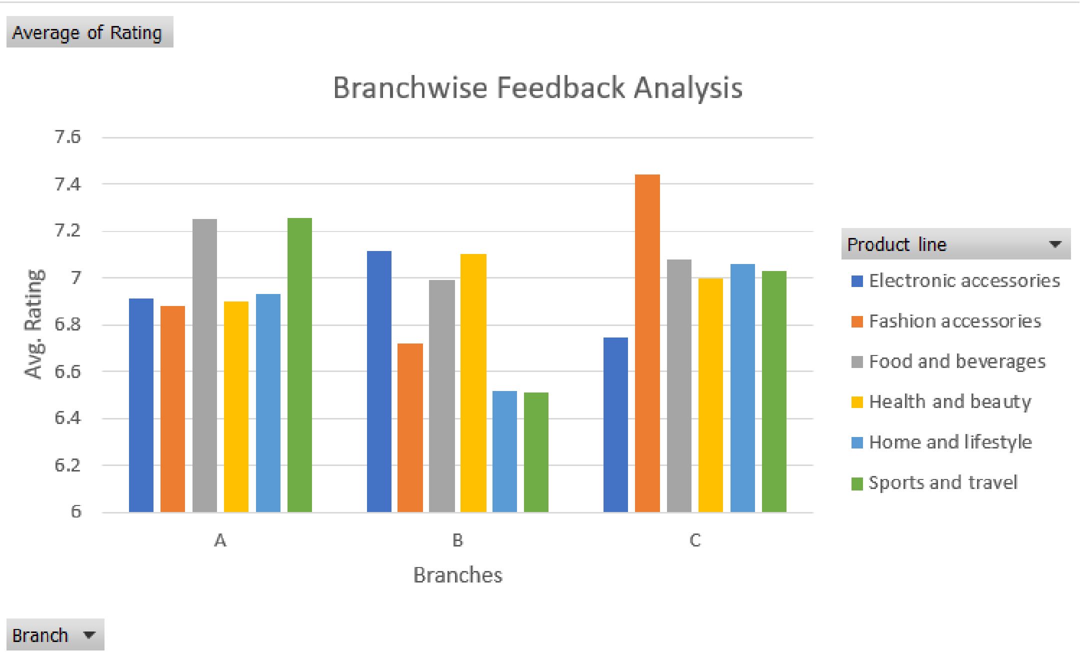


fig 5. Branch-wise Customer Feedback

Interpretations:

- The Home & lifestyle products need some improvement
- The Food & Beverages products need attention as the lead products (better inventory management and stock) as they have the most positive reviews
- Branch B has the harshest reviews so its services need to be reassessed.

Trend Analysis:

Monthly Sales for each branch were analysed to spot for any trends using a line chart:

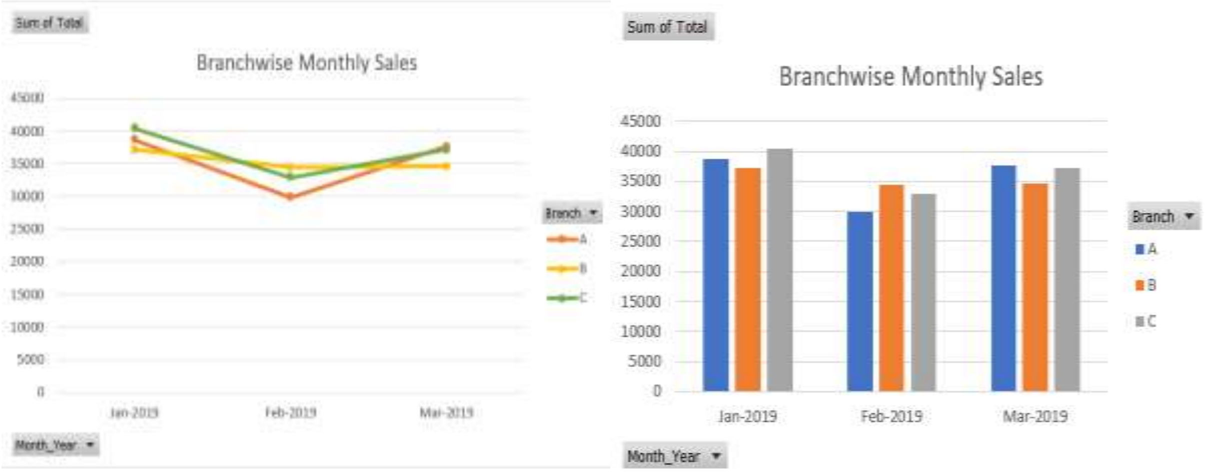


fig 6,7. Branch-wise Monthly Sales

Interpretations:

- Sales peaked in January indicating a possibility of post-holiday increase in demand.
- Sales saw a dip in February indicating seasonal drop followed by a recovery in March.
- All three branches are contributing similarly to overall revenue, but Branch C consistently recorded the highest monthly sales.

Daily sales were also considered to spot any other trends or patterns:

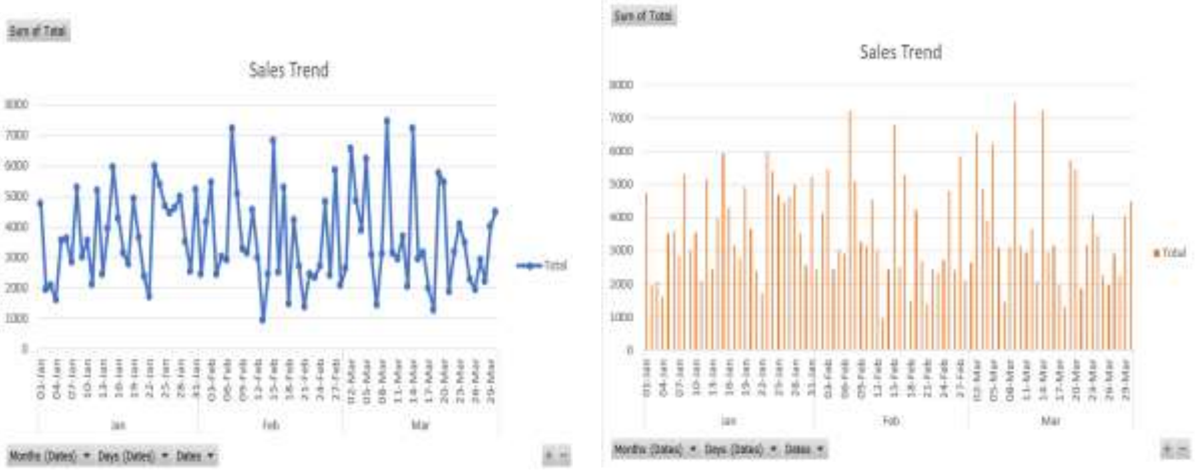


fig 8,9. Daily Sales Trend (All branches included)

Interpretations:

1. Sales peaked mid-month for all the 3 months which suggests that people demand and purchase supplies during the 2nd and 3rd weeks of the month.
2. Supply of goods and inventory management should be planned keeping this in mind.

Trends in profit were considered to see when the supermarket had more profits:

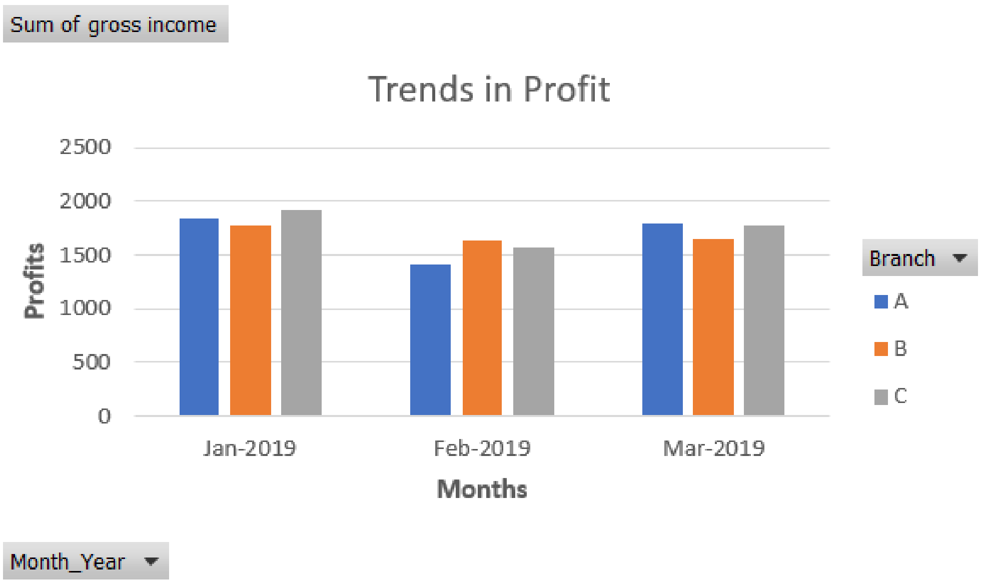


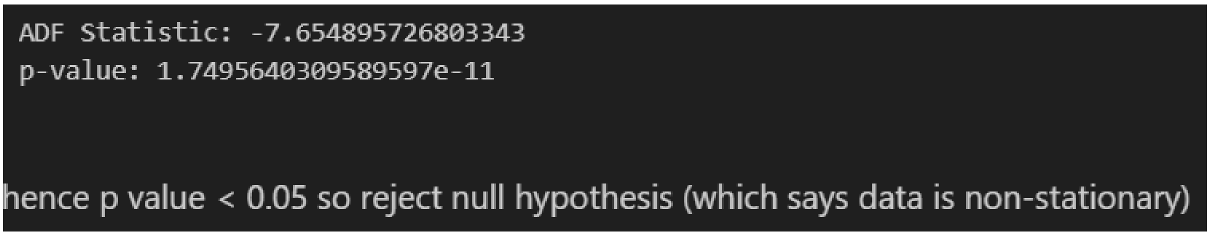
fig 10. Trends in Profit

Interpretations:

1. Profit trends mirror sales trends, implying that costs (e.g., inventory, staffing) are well-managed relative to revenue, but further cost optimization and profit maximization should be explored
2. Supply of goods and inventory management should be planned keeping this in mind.

Forecasting:

Historical sales data was used to predict future sales trends using **ARIMA** (AutoRegressive Integrated Moving Average) modeling; before that the data was checked with **Augmented Dickey-Fuller (ADF)** test to determine stationarity for ARIMA model which resulted in rejection of null hypothesis (that data is non-stationary). Then the 'p' and 'q' values were determined using the **ACF (AutoCorrelation Function)** and **PACF (Partial ACF)** plots:



```
ADF Statistic: -7.654895726803343
p-value: 1.7495640309589597e-11
```

hence p value < 0.05 so reject null hypothesis (which says data is non-stationary)

fig 11. ADF Test Results

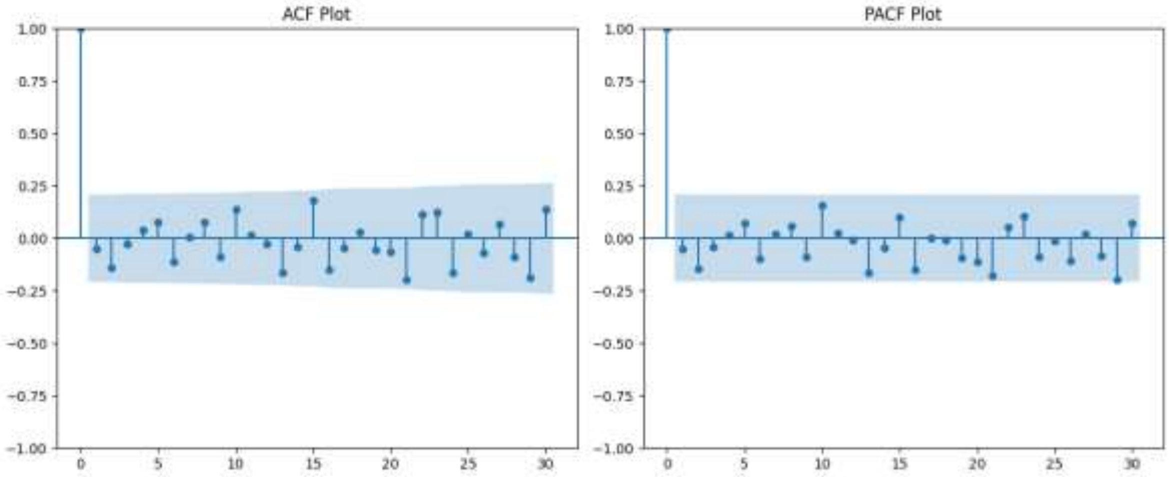


fig 12. ACF and PACF plots

Interpretations:

1. ACF plot: No significant autocorrelation after lag 1; suggests **q = 0 or 1**
2. PACF plot: Same behavior, only lag 1 stands out; suggests **p = 0 or 1**

Then the forecasting was done after fitting the ARIMA model and plotting the forecast using a line chart:

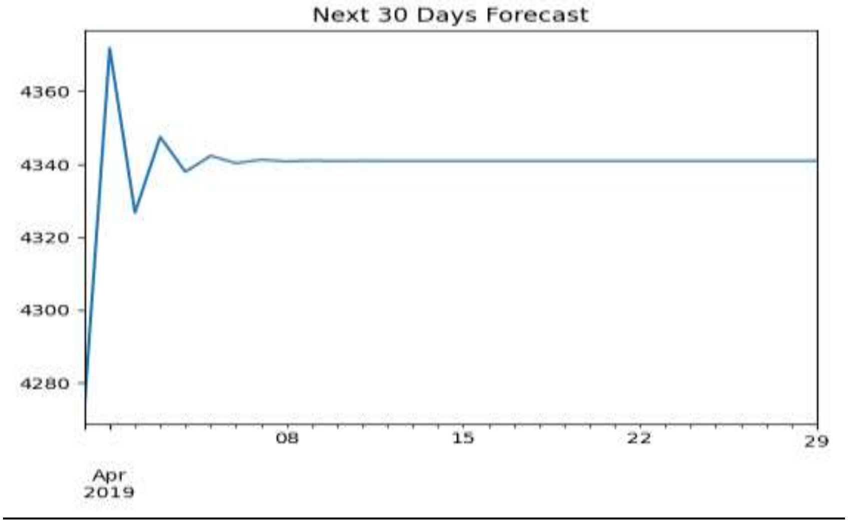


fig 13. ARIMA Forecast

Interpretations:

1. The first few days show noticeable fluctuations in forecasted sales.
2. From around April 8 onward, the forecast flattens out, indicating the model expects sales to remain steady; indicating the model did not detect strong seasonal or upward/downward trends beyond the short term

Seasonality seemed to affect the results of this ARIMA model so, SARIMA model was also fit and plotted:

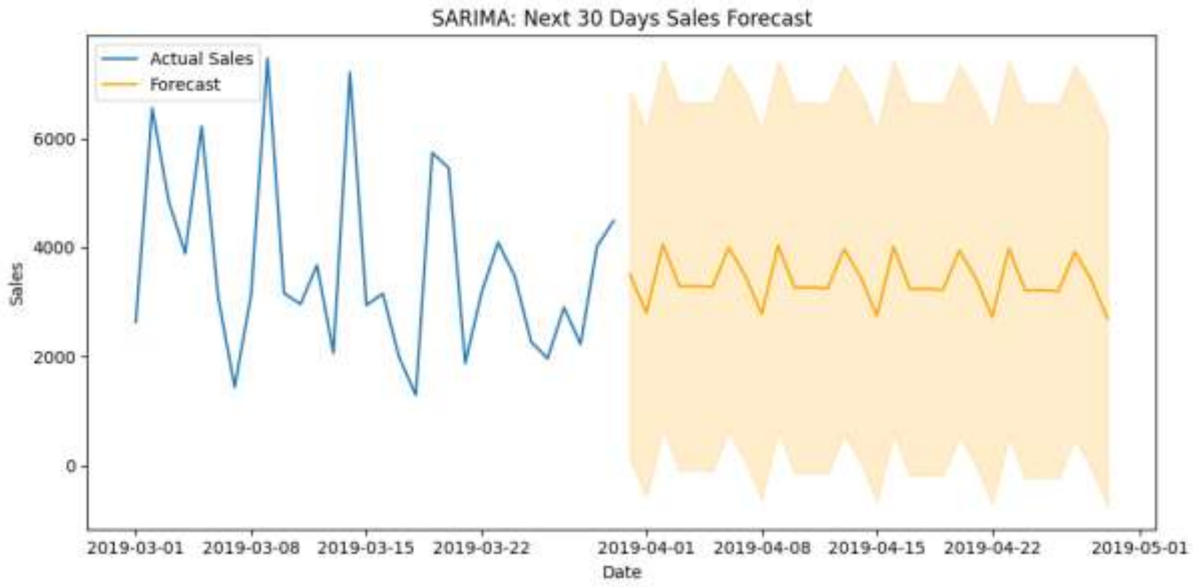


fig 14. SARIMA Forecast

Interpretations:

1. Displays a repeating weekly wave-like pattern, suggesting the model learned the weekly seasonality from March; The line is smoother than actual sales, as SARIMA focuses on underlying patterns and ignores random noise.
2. Shaded Region (confidence interval) is wide suggesting uncertainty, possibly due to volatility in the training data.

6 Interpretations of Results and Recommendations

Pareto Analysis Findings

Results:

- The 80/20 rule did **not** apply—revenue was **evenly distributed** across products and customer segments.
- No specific product or customer segment dominated sales, indicating **balanced purchasing behaviour**.

Recommendations:

- **Diversified Marketing:**
  Maintain broad promotional strategies rather than focusing on a select few products or customer groups.

- **Inventory Balance:**
  Ensure all product lines are well-stocked, as each contributes similarly to total revenue.
- **Category-Wide Promotions:**
  Run periodic promotions across multiple categories instead of select items to engage a wider customer base.
- **Customer Segmentation Exploration:**
  Conduct deeper segmentation (e.g., demographics, shopping frequency, basket size) to uncover micro-trends for targeted offerings.

Customer Feedback Analysis Findings

Results:

- **Home & Lifestyle** received the **lowest customer satisfaction ratings**, signaling areas for improvement.
- **Food & Beverages** received the **highest ratings**, indicating strong customer satisfaction.
- **Branch B** had the most negative feedback so it must be reassessed for product quality and employee training.

Recommendations:

- **Improve Home & Lifestyle:**
  - Investigate product quality, pricing strategies, and visual merchandising.
  - Collect direct feedback via in-store or digital surveys.
  - Consider rotating stock or introducing new, high-demand lifestyle products.
- **Leverage Food & Beverages:**
  - Use positive customer sentiment as a benchmark for other categories.
  - Maintain supplier relationships to ensure consistent availability and quality.
  - Highlight top-performing food items in promotions or loyalty programs.
- **Cross-Category Insights:**
  - Analyze what drives satisfaction in Food & Beverages (e.g., freshness, pricing, packaging) and apply similar strategies in underperforming departments.
- **Staff Training and Engagement:**
  - Train staff in the Home & Lifestyle section to better assist customers and create a more engaging shopping experience.

Trend Analysis Findings

Results:

- **Monthly Trends:**
  - Peak sales in **January** (post-holiday demand), drop in **February**, slight recovery in **March**.
- **Daily Trends:**
  - Sales peaked consistently in the **2nd and 3rd weeks** of each month.
- **Trends in Profit:**
  - Sales & Profit Trends Are **Highly Correlated**, both peak mid-month.

Recommendations:

- **Seasonal Planning:**
  - Increase inventory and staffing in **January** to capture high demand.
  - Use promotions and loyalty incentives to stimulate demand in **February**.
- **Mid-Month Focus:**
  - Align restocking and marketing efforts with the **2nd and 3rd weeks** of each month.
  - Offer mid-month exclusive deals to encourage repeat visits.
- **Calendar-Based Campaigns:**
  - Align promotions with local events, pay cycles, and holidays to match shopping patterns.
- **Dynamic Staffing:**
  - Implement a flexible staffing model to adjust employee schedules based on forecasted peak days and hours.

Forecasting Findings

Results:

- The SARIMA model captured a clear **weekly seasonality**, predicting consistent sales cycles throughout April.
- Forecasted sales show a **stable trend** with moderate fluctuations, indicating **predictable demand patterns**.

- Confidence intervals reflect **uncertainty due to past volatility**, but still suggest manageable operational planning.

Recommendations:

- **Weekly Demand Planning:**
  - Align staffing, inventory, and promotions around the forecasted weekly sales cycles to optimize efficiency and resource use.
- **Monitor Forecast Accuracy:**
  - Track actual sales against the forecast to identify anomalies (e.g., due to holidays or campaigns) and retrain the model periodically.
- **Refine Seasonality Detection:**
  - Incorporate longer time periods or external factors (e.g., holidays, events) to improve seasonal accuracy.
- **Scenario Planning:**
  - Use confidence intervals to simulate best- and worst-case sales scenarios for better financial forecasting and risk management.

7 Overall Strategic Summary

1. **Balanced Revenue Distribution:**
   - Maintain consistent quality and stock across all categories.
   - Avoid over-dependence on limited products or customer types.
   - Explore micro-segmentation for future personalization:
     - Use purchase history to tailor promotions (e.g., discounts on frequently bought items).
2. **Feedback-Driven Improvement:**
   - Prioritize enhancements in **Home & Lifestyle**.
     - Redesign store layouts to improve product visibility and accessibility
     - Partner with local artisans or brands to offer exclusive Home & Lifestyle products
   - Apply winning strategies from **Food & Beverages** across departments.
     - Replicate successful promotions (e.g., "Buy One, Get One" deals) in underperforming sections.

    - Train staff in cross-selling techniques to boost complementary purchases (e.g., promoting kitchenware with groceries)

- Use feedback loops to continually evolve the product mix and store experience.

- Train the staff (especially in the Home & Lifestyle section) to better assist customers and create a more engaging shopping experience.

3. **Demand Forecasting:**

    - Use monthly and daily sales trends to drive data-backed inventory and staffing decisions.

    - Design marketing and pricing strategies to smooth out seasonal slumps and capitalize on peak periods (for example: Align restocking and marketing efforts with the **2nd and 3rd weeks** of each month).

    - Implement flexible staffing model to adjust employee schedules based on forecasted peak days and hours.

4. **Enhanced Data Collection:**

    - Invest in improved data collection systems (e.g., POS upgrades, loyalty programs, customer surveys).
    - Retain and store long-term sales and customer behavior data for trend analysis and predictive modeling.
    - Encourage digital receipts and online engagement to gather richer data on customer habits.
    - Establish better data management protocols and maintain an extensive historical database to support deeper forecasting and strategic planning.

5. **Profit Margin Optimization**:

- Analyse pricing strategies and supplier contracts to identify margin improvement opportunities.
- Introduce private label or high-margin products to boost gross margin percentage.
- Use category-level margin data to optimize product placement, promotion, and markdown strategies.