Smart Analytics for a Vegetable Seller

The End-Term report for the BDM capstone Project

Submitted by

Name: Anuj Krishna S

Roll number: 21F2000852

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1 Executive Summary

Introduction:

Mr. Muthuraj A, a long-time resident of Chennai, has been operating a roadside vegetable shop in Ashok Nagar for over 35 years. With deep roots in the local community and strong customer relationships, his shop is a familiar and trusted presence in the neighborhood. As a sole proprietor, Mr. Muthuraj has built his business through years of hands-on experience and personal rapport with customers, relying heavily on intuition and routine rather than formal business practices. His dedication and consistency have earned him a loyal customer base and a steady daily flow of foot traffic.

Challenges:

Despite his experience and reputation, Mr. Muthuraj's business has been facing two major challenges in recent years—stagnant sales and high levels of vegetable wastage. Sales have remained flat over time, limiting revenue growth and reducing overall profitability. At the same time, the perishable nature of vegetables has led to significant losses due to unsold and spoiled stock. The absence of structured records for sales, stock, and procurement has made it difficult to understand demand trends, manage inventory effectively, or adjust purchasing practices based on actual needs. These operational inefficiencies have created a cycle of oversupply and wastage, further straining business margins.

Project Approach:

To address these challenges, this project adopts a data-driven approach focused on structured data collection and analysis. Daily records of customer purchases, procurement prices, stock levels, and remaining inventory are maintained to build a comprehensive dataset. Using this data, the project applies analytical techniques to identify demand patterns, optimize stock procurement, and introduce waste reduction strategies. The goal is to improve inventory control, reduce spoilage, and support better pricing and sales strategies. By introducing these data-backed practices, the project aims to transform Mr. Muthuraj's operations into a more sustainable and profitable business model, while also offering a replicable framework for similar small-scale vendors.

2 Detailed Explanation of Analysis Process/Method

2.1 Data Cleaning and Preprocessing:

I have used the below pandas code for the data cleaning process.

```
# Convert 'Date' to datetime
for df in [market_df, sales_df, customer_df]:
    df['Date'] = pd.to_datetime(df['Date'])

# Clean 'Qty(KG)' column
market_df['Qty(KG)'] = market_df['Qty(KG)'].str.replace("Kg", "").astype(float)

# Merge Sales + Market
merged_df = pd.merge(sales_df, market_df, on=["Date", "Product_Id"], suffixes=('_Sales', '_Market'))

# Merge with Product_id
merged_df = pd.merge(merged_df, product_df, on="Product_Id", how="left")
```

This code performs data cleaning and preprocessing on an Excel file with multiple sheets related to a market sales project. It begins by uploading the Excel file and reading four sheets: ‘Market’, ‘Sales’, ‘Product_id’, and ‘Customer’. It converts the 'Date' column in the relevant data frames to datetime format for consistency. In the Market sheet, it cleans the 'Qty(KG)' column by removing the "Kg" suffix and converting the values to float. It then merges the *Sales* and Market data frames on 'Date' and 'Product_Id', followed by a merge with the Product_id data frame to enrich the dataset with product details. The combined dataset is stored as ‘merged_df’.

2.2 Price vs Demand Analysis:

To perform a Price vs. Demand Analysis,

- **Check Price Elasticity:** This involves calculating how the quantity sold changes in response to price changes.

$$\text{Price Elasticity of Demand} = \frac{\%\ \text{Change in Quantity Demanded}}{\%\ \text{Change in Price}}$$

The interpretations for the price elasticity od demand (PED):

- If **PED > 1**, then it is elastic (quantity changes more than price)
- If **PED < 1**, then it is inelastic (quantity changes less than price)
- If **PED = 1**, then it is unit elastic
- If **PED = 0**, then it is perfectly inelastic
- If **PED =** ∞, then it is perfectly elastic

- **Use Scatter Plots and Correlation Analysis:** The document includes a scatter plot of % Margin vs. Profit and performs a linear regression. A similar approach could be used to plot Price vs. Quantity Sold and analyze the correlation.

2.3 Profitability Analysis:

- Vegetable Cost = Total Cost Incurred while procuring all vegetables for the day
- Need to compute the revenue, profit and the margin% for the profit for all the products.
- The formula for the revenue for the products:

$$\boldsymbol{Revenue = Vegetable\ price * Quantity}$$

The profit is calculated by:

$$\boldsymbol{Profit = Revenue - Vegetable\ Cost}$$

- Most Profitable Products: Margin Analysis provides insights into product profitability:

$$\boldsymbol{Margin = CP - SP}$$

2.4 Time Series Analysis:

Time series analysis is a statistical method used to analyse data points collected over time, aiming to identify patterns, trends, and seasonality. It helps in forecasting future values and understanding the underlying structure of the data. The time series analysis involves the following steps:

- **Aggregate sales data:** The sales data is aggregated by date and product to provide a summarized view of sales performance over time.
- **Visualize trends:**
    - **Line plots:** Daily sales are visualized using line plots to illustrate the fluctuations and patterns in sales activity.
    - **Moving average:** A moving average is applied to smooth out short-term fluctuations and highlight longer-term trends in the sales data.

There are two statistical models that can be used when working with certain pattern. They are,

2.4.1 ARIMA – Autoregressive Integrated Moving Average:

It is used for non-seasonal time series data. It combines three parts:

AR (Auto Regressive) predicts the value based on past values (lags). I (Integrated) is used for differencing the data to make it stationary (removes trends). MA (Moving Average) uses past forecast errors to improve the prediction.

2.4.2 SARIMA – Seasonal ARIMA:

SARIMA extends ARIMA to handle seasonal data. We can use SARIMA, when data patterns repeat at regular intervals (like weekly or monthly). It adds seasonal components to the ARIMA model.

2.5 Customer Pattern Clustering:

K-means clustering: K-means clustering was applied to segment customers based on their spending and purchase quantity. This analysis identified three distinct customer groups:

- **Cluster 1: High Spenders, Large Quantity**
    - Customers in this cluster exhibit high spending and purchase large quantities.
- **Cluster 2: Low Spenders**
    - This cluster comprises customers who spend relatively less.

- **Cluster 0: Bulk Buyers at Low Spend**
    - Customers in this segment purchase in bulk but at a lower spending level.

2.6 ABC Analysis:

ABC analysis is a method of categorizing inventory items based on their value or importance to a business. It's a way to prioritize inventory management efforts.

- **Category A:** These are the high-value items that contribute the most to the overall inventory value. They typically represent a small percentage of the total inventory items but a large percentage of the total sales or usage value.
- **Category B:** These items have a moderate value. They fall between Category A and Category C in terms of importance.
- **Category C:** These are the low-value items that contribute the least to the overall inventory value. They usually represent a large percentage of the total inventory items but a small percentage of the total sales or usage value.

2.6.1 Role in Stock Management:

ABC analysis plays a crucial role in stock management by enabling businesses to:

- **Prioritize Control:** It allows businesses to focus their attention and resources on the items that have the greatest impact on profitability.
- **Optimize Resource Allocation:** By understanding the value of different inventory items, businesses can allocate resources more effectively. For example, more time and effort can be spent on managing Category A items compared to Category C items.
- **Improve Inventory Control Policies:** Different control policies can be applied to different categories of inventory. For instance, Category A items may require more frequent monitoring, tighter security, and more accurate forecasting compared to Category C items.
- **Reduce Costs:** By optimizing inventory control, businesses can reduce carrying costs, ordering costs, and the risk of stockouts.

2.7 Stock Optimization Strategy:

The stock optimization strategy involves the following steps:

- **Daily Average Sales (Moving Average):** Calculate a 7-day rolling average to estimate the quantity of each product sold daily. This approach captures recent demand trends.
  - Example: If 6.2 KG of tomatoes were sold daily on average over the past week, this figure represents the starting point for the demand estimate.
- **Reorder Point (ROP):** Determine the Reorder Point (ROP) using the following formula:

$$\boldsymbol{ROP = Average\ Daily\ Sales * Lead\ Time}$$

  - Where Lead Time is the time required for restocking (in this case, 2 days).
  - Example: For tomatoes, the ROP would be 6.2 KG/day × 2 days = 12.4 KG.
- **Perishability Classification:** Classify each product based on its perishability:
  - High Perishability: Products that spoil quickly.
  - Medium Perishability.
  - Low Perishability: Products with a stable shelf life.
- **Adjusted Optimal Stock Level:** Adjust the order quantity based on the perishability classification:
  - High Perishability: Order 80% of the ROP.
  - Medium Perishability: Order 100% of the ROP.
  - Low Perishability: Order 120% of the ROP.

This adjustment strategy aims to minimize waste for high-risk items while ensuring an adequate supply of low-risk items.

3 Results And Findings:

Based on the analysis of the project, the results and findings are explained with graphical representations.

3.1 Profit Trend Overview:

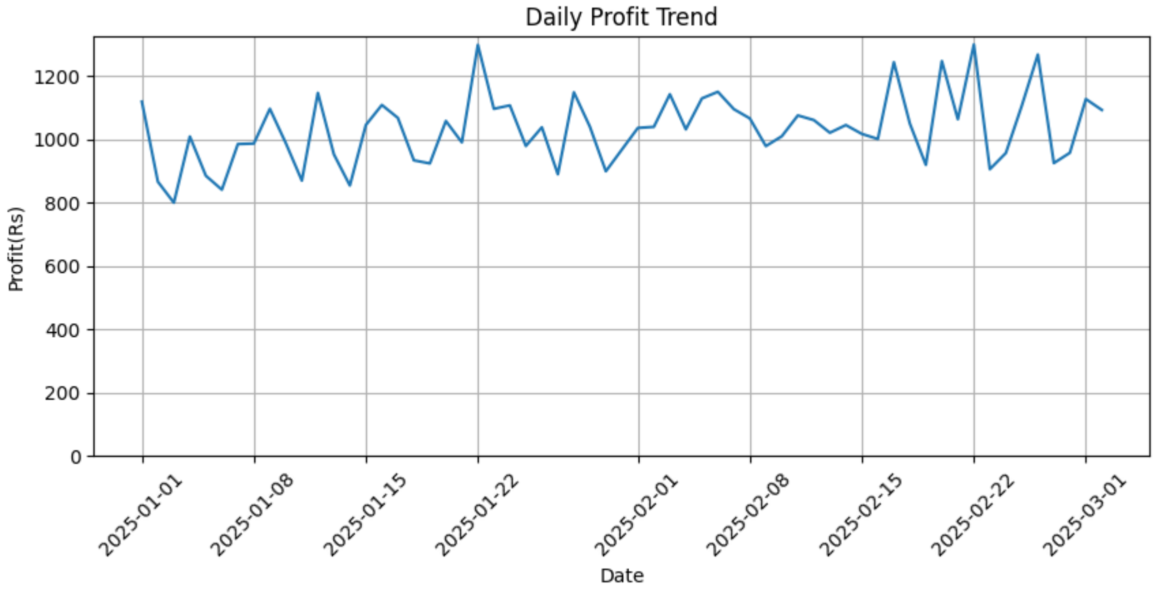

Fig 1. Daily Profit Trend

- The daily profit exhibits regular fluctuations over the two-month period (January–February 2025).
- Profits generally range between ₹900 and ₹1200, with a few days showing exceptional spikes above ₹1200.
- The overall pattern does not indicate a consistent upward or downward trend, suggesting stable yet variable daily profits.
- **High-Profit Spikes**: Occur notably around January 21 and February 21, reaching peak values.
- **Low-Profit Dips**: A few dips closer to ₹800 are seen in early January.
- **Stable Period**: A more stable profit pattern is observed during early to mid-February.

3.1.1 Total Profit Generated by the products:

The plot describes the overall profit for all the products over the time duration.

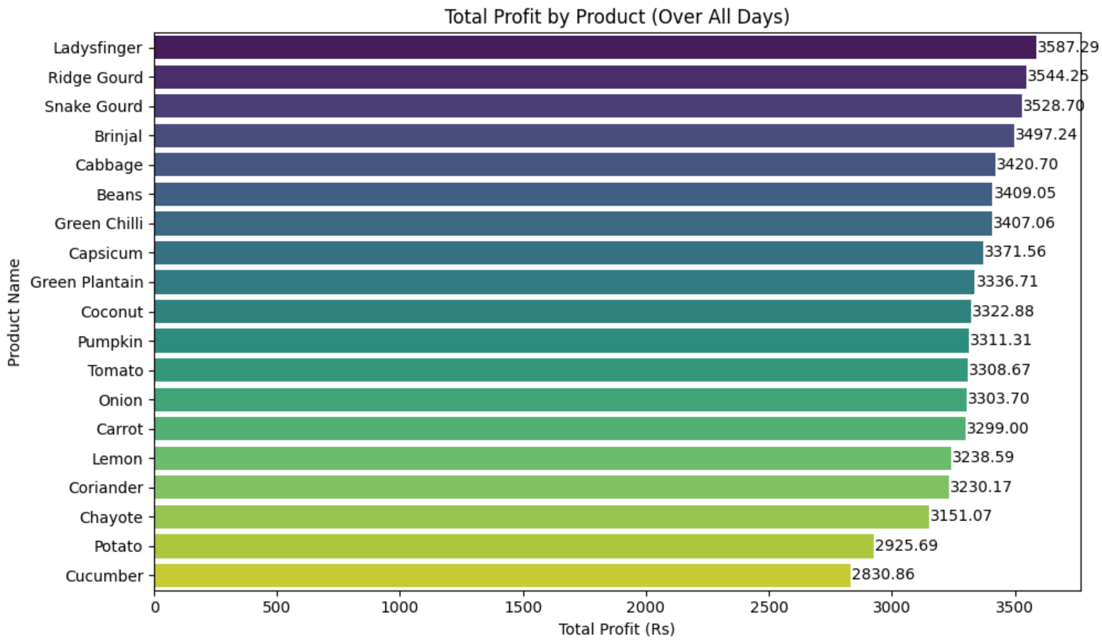

Fig 2. Total profit for all the products

Analysis:

- The chart provides a clear comparison of the overall profitability of each product.
- 'Lady's Finger' generated the highest total profit (Rs 3587.29), followed closely by 'Ridge Gourd' (Rs 3544.25) and 'Snake Gourd' (Rs 3528.70). These three products are the most profitable in the given dataset.
- 'Cucumber' generated the lowest total profit (Rs 2830.86), and 'Potato' is also among the lower-performing products in terms of total profit (Rs 2925.69).
- The remaining products fall between these extremes, showing varying levels of profitability.

3.2 Profit and Profit Margin Trend Over Time:

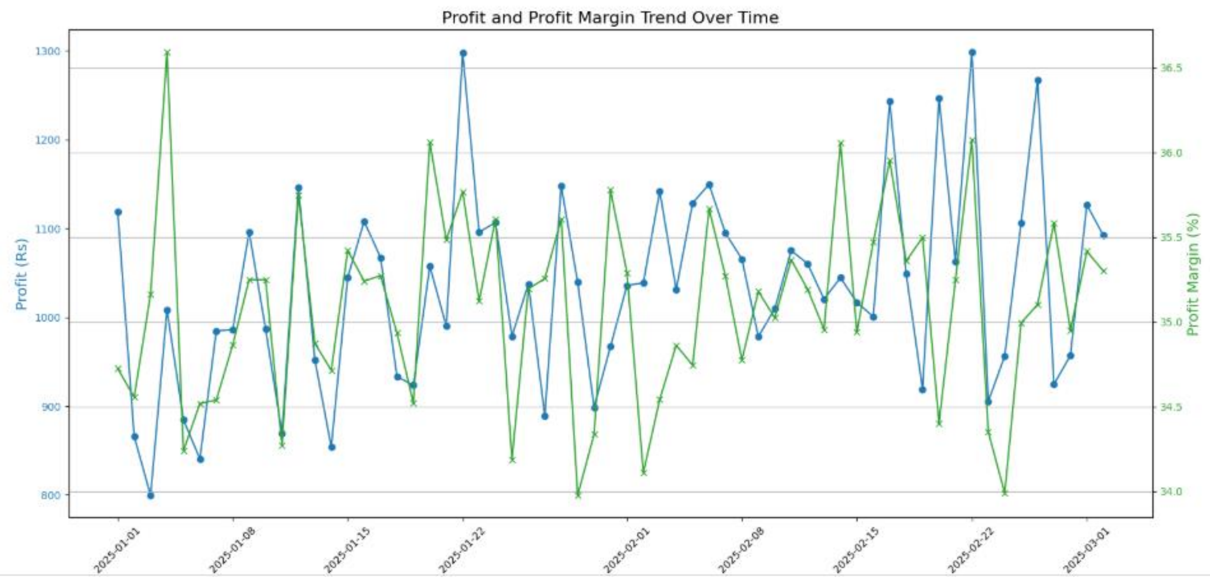

Fig 3. Profit and Profit Margin%

The plot illustrates the fluctuations in profit (Rs) and profit margin (%) over a period.

- **Profit Trend:** Profit exhibits significant volatility, with sharp peaks and troughs throughout the period. This indicates that profit is subject to substantial fluctuations.
- **Profit Margin Trend:** The profit margin also fluctuates, but within a narrower range compared to profit. This suggests that while profit varies considerably, the profit margin remains relatively more stable.
- **Relationship:** There is no consistent correlation between profit and profit margin. High profit values do not always correspond to high profit margins, and vice versa. This implies that factors other than profit margin, such as sales volume and costs, significantly influence profit.

Statistical Analysis of Profit Margin:

- Mean Profit Margin: 35.10%
- Median Profit Margin: 35.17%
- Standard Deviation of Profit Margin: 2.90%
- The mean and median profit margin are very close, indicating a roughly symmetric distribution of profit margins.

- The standard deviation of 2.90% suggests that the profit margin typically deviates from the average by about 2.90%. This indicates a moderate level of variability in profit margins.

3.3 Price Elasticity Heatmap Analysis:

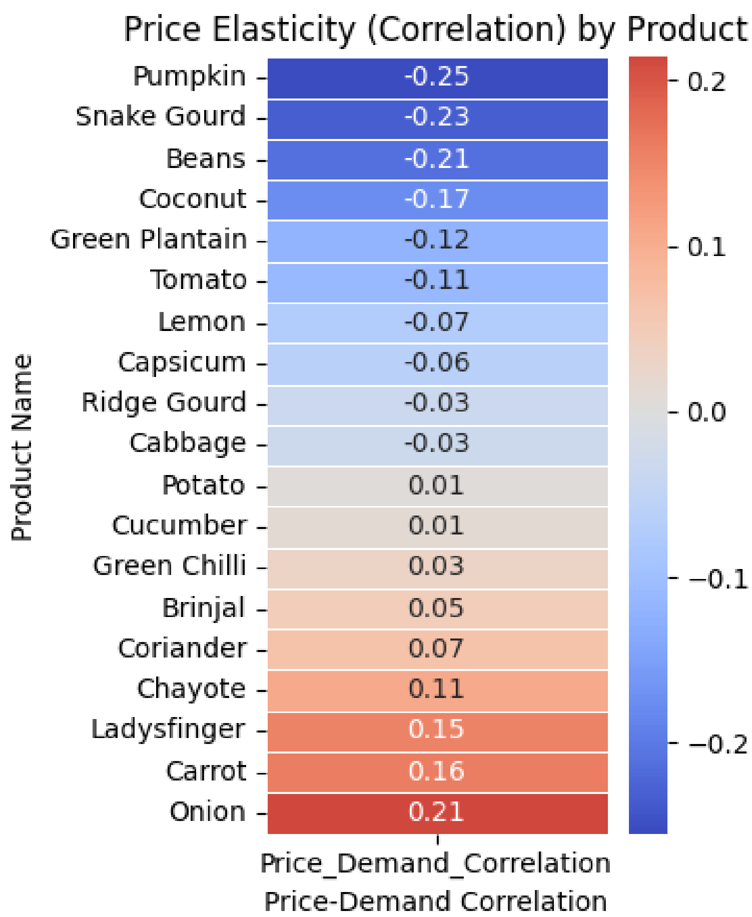

Fig 4. Price Demand Correlation

Visual Representation:

- The heatmap displays the price elasticity of demand for various products.
- The colour scale ranges from blue (negative correlation, elastic demand) to red (positive correlation, inelastic demand).
- The heatmap shows the correlation between price and demand for different products.

Key Observations:

- Products like 'Pumpkin', 'Snake Gourd' and 'Beans' show a more negative correlation, suggesting that their demand is more sensitive to price changes. These products have a more elastic demand.
- Products like 'Onion', 'Carrot', and 'Lady's Finger' have a positive correlation between price and demand. This indicates that as the price of these products increases, the demand also tends to increase, though this is counterintuitive.
- Products like 'Cabbage', 'Ridge Gourd', 'Potato', and 'Cucumber' show a correlation close to zero, indicating that their demand is relatively inelastic, meaning it doesn't change much with price fluctuations.

3.4 Time Series Analysis:

Time series analysis is a method used to understand and analyse data that is collected over time at regular intervals (daily, weekly, monthly, etc.). It helps in identifying patterns such as trends**,** seasonality, and cycles, and is often used to make forecasts or predictions about future values.

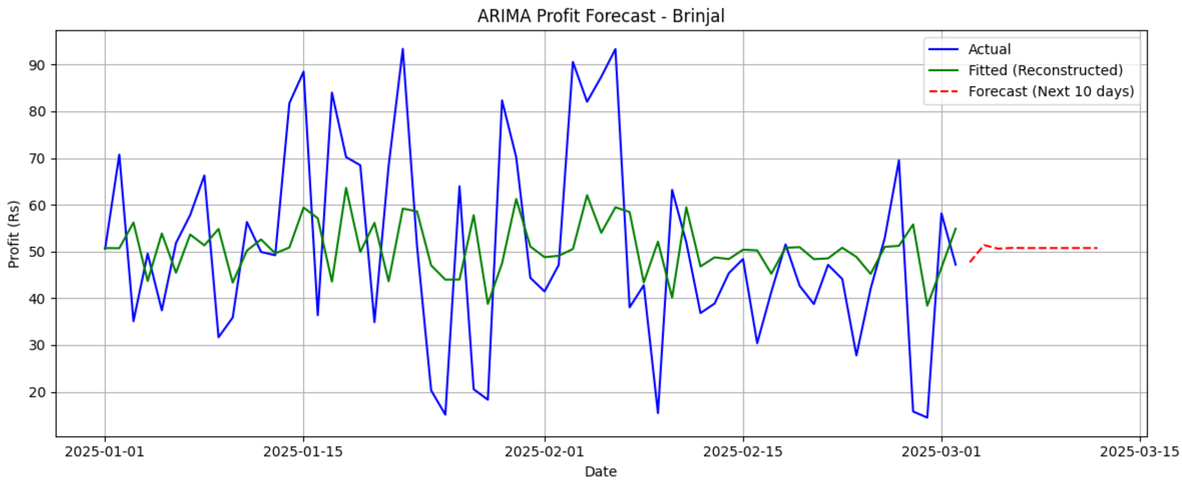

Fig 5. Profit Vs Brinjal

This graph shows the ARIMA profit forecast for Brinjal from January to mid-March 2025. The blue line represents the actual daily profit, which shows high fluctuations over time. The green line is the fitted or reconstructed profit values using the ARIMA model, closely following the overall trend of the actual data. The red dashed line on the right-side forecasts the profit for the next 10 days after March 1, 2025. According to the forecast, the profit is expected to remain fairly stable around Rs 50**.**

We use SARIMA and check for the seasonality present for the product over the time.

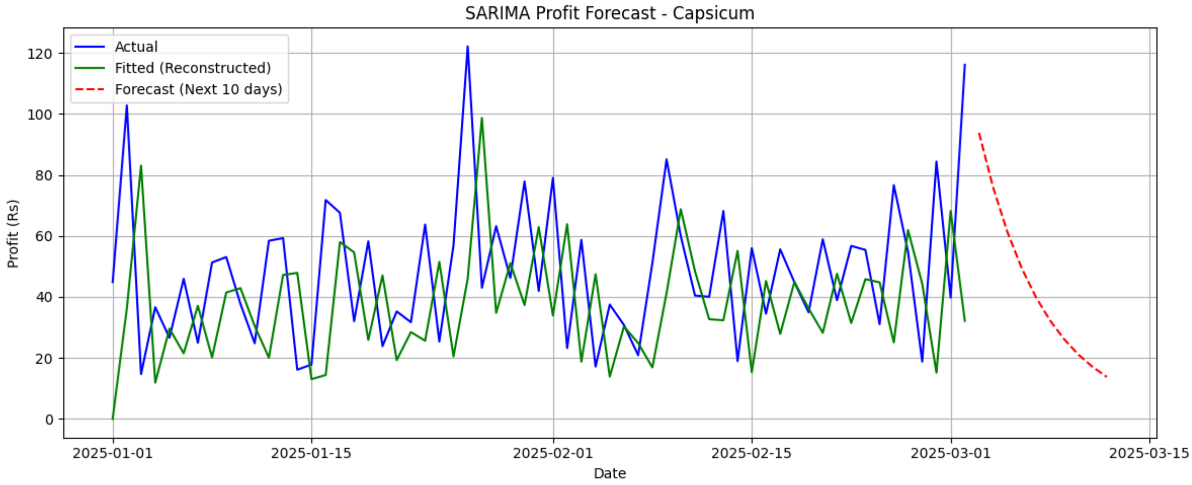

Fig 6. Profit Vs Capsicum

This graph shows the SARIMA profit forecast for Capsicum from January to mid-March 2025. The blue line represents the actual profit recorded each day. The green line shows the fitted (reconstructed) values from the SARIMA model, which closely follow the actual data. The red dashed line on the right side shows the forecasted profit for the next 10 days after March 1, 2025. According to the forecast, the profit is expected to decrease gradually.

3.5 Customer Pattern Clustering:

K-means clustering was applied to segment customers based on their spending and purchase quantity.

Fig 6. Spend Vs Quantity Purchased

This analysis identified three distinct customer groups:

- **Cluster 1: High Spenders, Large Quantity**
    - Customers in this cluster exhibit high spending and purchase large quantities.
- **Cluster 2: Low Spenders**
    - This cluster comprises customers who spend relatively less.
- **Cluster 0: Bulk Buyers at Medium Spend**
    - Customers in this segment purchase in bulk but at a medium spending level.

3.5.1 Cluster Summary:

	Cluster	Avg_Spend	Min_Spend	Max_Spend	Avg_Quantity	Min_Quantity	Max_Quantity	Customer_Count
0	0	4432.490000	4124.170000	4803.730000	110.450000	103.000000	115.750000	16
1	1	5110.920000	4873.000000	5417.610000	125.160000	118.500000	132.250000	8
2	2	3810.900000	3700.910000	3933.480000	98.960000	96.000000	105.500000	6

Table 1. Cluster Summary

Cluster 0: This segment comprises 16 customers with an average spend of Rs 4432.49, ranging from Rs 4124.17 to Rs 4803.73, and an average purchase quantity of 110.45 units, between 103 and 115.75 units. These customers represent a substantial portion of the customer base and exhibit consistent spending and purchasing patterns, with relatively low variability within the group.

Cluster 1: This cluster contains 8 customers demonstrating the highest spending and purchasing behaviour, with an average spend of Rs 5110.92 (between Rs 4873 and Rs 5417.61) and an average quantity purchased of 125.16 units (between 118.5 and 132.25 units). This group is characterized by the highest engagement.

Cluster 2: This segment includes 6 customers with the lowest spending and purchasing levels, averaging Rs 3810.90 per customer (between Rs 3700.91 and Rs 3933.48) and 98.96 units per customer (between 96 and 105.5 units). This cluster represents the smallest customer base and has the lowest activity.

3.6 ABC Analysis:

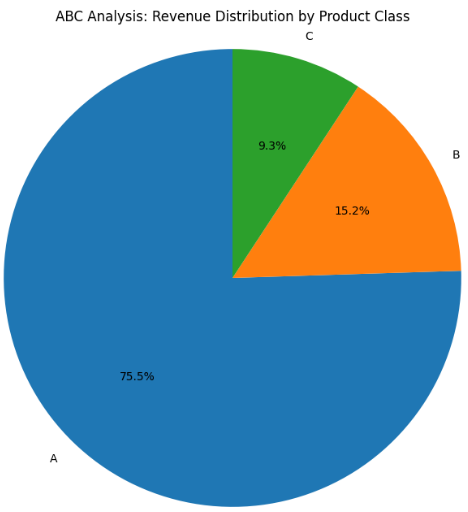

Fig 7. ABC Analysis with Revenue

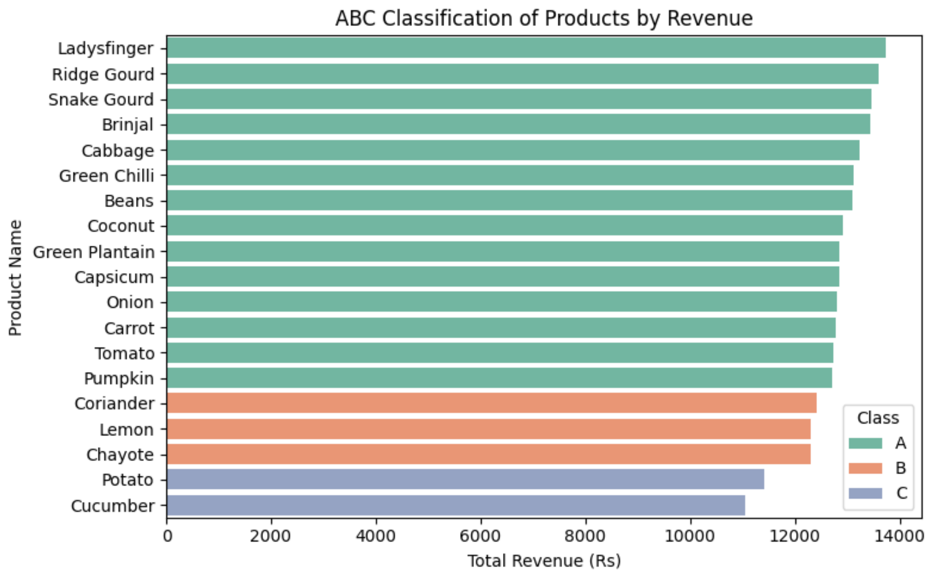

Fig 8. Revenue Vs Product

The plot illustrates the ABC classification of products based on their total revenue. The products are categorized into three classes: A, B, and C, with Class A representing the highest revenue-generating products and Class C the lowest.

- **Class A:** This class accounts for the largest share of revenue, with 75.5%. It includes products such as Lady's Finger, Ridge Gourd, and Snake Gourd, which generate the highest revenue. These products are critical to the business's overall revenue and should be given the highest priority in terms of inventory management and sales strategies.
- **Class B:** This class contributes 15.2% of the total revenue. This category comprises products like Coriander, Lemon, and Chayote. These products generate moderate revenue.
- **Class C:** This class represents the smallest portion of revenue, with only 9.3%. This includes products like Potato and Cucumber, which generate the lowest revenue. While these products are still important, they require less attention and resources compared to Class A products.

Key Observations:

- A large number of products fall into Class A, indicating that revenue is concentrated across these top-performing items.
- There are fewer products in Class C, suggesting that only a small portion of the product range contributes minimally to the overall revenue.

Implications:

- The business should focus on maintaining a high level of availability for Class A products to maximize revenue.
- Inventory control and demand forecasting should be most stringent for Class A products.
- For Class C products, a simpler inventory management approach may be sufficient.

3.7 Effective Stock Management Techniques:

Analysis of Top 10 High-Wastage (Perishable) Vegetables:

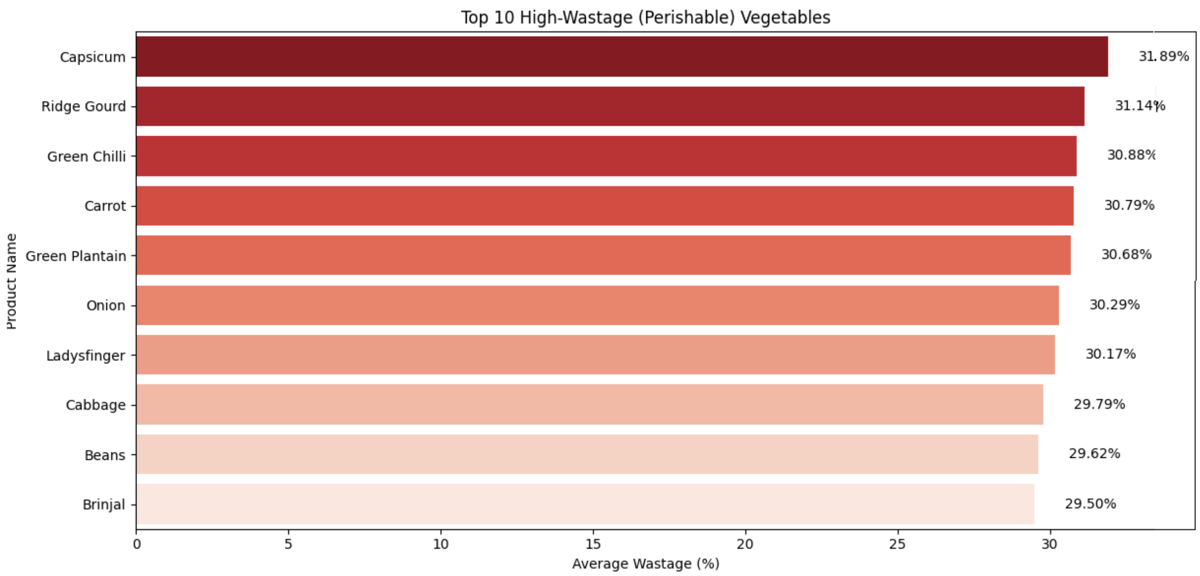

Fig 9. Wastage Vs Product

The bar chart illustrates the average wastage percentage for the top 10 perishable vegetables.

Key Findings:

- Capsicum has the highest wastage at 31.89%, closely followed by Ridge Gourd at 31.14% and Green Chilli at 30.88%.
- Brinjal shows the lowest wastage among the top 10, with 29.50%.

- The wastage percentage for all the vegetables in the chart is quite high, ranging from approximately 29.50% to 31.89%. This indicates a significant issue with perishability across these vegetables.

```python
# Check correlation
correlation = merged_df["Stock_Purchased(KG)"].corr(merged_df["Wastage(KG)"])
print(f"Correlation between Stock Purchased and Wastage: {correlation:.2f}")
```

```
Correlation between Stock Purchased and Wastage: 0.63
```

A correlation coefficient of +0.63 was observed between the quantity of stock purchased and wastage incurred. This moderately strong positive relationship suggests that overstocking is a key driver of wastage, especially in the case of perishable or slow-moving products. It reinforces the need for improved demand forecasting and dynamic stock planning to prevent unnecessary losses and optimize resource utilization.

3.7.1 Stock Optimization Strategy:

The goal of the strategy is to prevent over-purchasing, especially for perishable vegetables, by using a smart, dynamic method to calculate how much stock should actually be ordered each day. After computing the key metrics such as "Daily Average Sales (Moving Average), Reorder Point (ROP), Perishability Classification, Adjusted Optimal Stock Level".

a) Daily Moving Average Sales:

```python
# Compute 7-day moving average of Stock Sold
merged_df["Daily_Avg_Sales"] = merged_df.groupby("Product_Name_Sales")["Stock_Sold(KG)"].transform(lambda x: x.rolling(window=7, min_periods=1).mean())
merged_df["Daily_Avg_Sales"]
```

This technique is used to smooth out daily sales data fluctuations and identify trends. A moving average calculates the average sales over a specific period of 7 days by continuously updating the average as new data becomes available. It helps in forecasting and understanding the underlying sales pattern without being swayed by single-day spikes or dips.

b) Reorder Point (ROP):

```python
# Reorder Point
lead_time = 2
merged_df["Reorder_Point"] = merged_df["Daily_Avg_Sales"] * lead_time
merged_df['Reorder_Point']
```

The Reorder Point is the minimum level of inventory that triggers a new order for a specific item. Its purpose is to ensure that you don't run out of stock before the next order arrives. In essence, for every product, the ROP is calculated using "Stock Sold" over the past 7 days and computes the average, storing this average in the "Daily_Avg_Sales" column.

c) Perishability Classification:

```python
# Perishability
# Average wastage per product
avg_wastage = merged_df.groupby("Product_Name_Sales")["Wastage_%"].mean()

# Add category: High, Medium, Low Perishability
def classify_perishability(w):
    if w > 30: # If wastage greater than 30%
        return "High"
    elif w > 27:
        return "Medium"
    else:
        return "Low"

merged_df["Perishability_Risk"] = merged_df["Product_Name_Sales"].map(avg_wastage).apply(classify_perishability)
merged_df["Perishability_Risk"].unique()
```

```
array(['Low', 'Medium', 'High'], dtype=object)
```

This categorizes products based on how quickly they degrade or expire. This classification is crucial for inventory management as it directly impacts storage, handling, pricing, and waste management. The code calculates the average spoilage for each product and then categorizes each product as "High", "Medium", or "Low" perishability based on predefined thresholds of average wastage. This classification is vital for inventory management strategies.

d) Adjusted Optimal Stock Level:

```python
# Adjusted Optimal Stock level
def adjust_stock(row):
    if row["Perishability_Risk"] == "High":
        return 0.8 * row["Reorder_Point"]
    elif row["Perishability_Risk"] == "Medium":
        return row["Reorder_Point"]
    else:
        return 1.2 * row["Reorder_Point"]

merged_df["Optimal_Stock_Level"] = merged_df.apply(adjust_stock, axis=1)
merged_df["Optimal_Stock_Level"]
```

The "Adjusted Optimal Stock Level" for each product is calculated, taking into account its "Perishability_Risk". It aims to manage inventory more effectively by stocking less of highly perishable items and potentially more of low-perishability items. In essence, this code adjusts the target stock level for each product based on how quickly it might spoil. It reduces the stock for easily perishable items and allows for more stock for items that last longer, optimizing inventory holding to reduce waste and potentially improve availability.

After this, we estimate the "Avoidable Overstock" of all the products.

```
# Only for rows where stock purchased exceeds optimal, Estimate avoidable wastage by checking:
avoidable_wastage = merged_df[merged_df["Stock_Difference"] > 0]["Stock_Difference"].sum()
print(f"Estimated Avoidable Overstock (KG): {avoidable_wastage:.2f}")
```

```
Estimated Avoidable Overstock (KG): 173.48
```

Based on our optimized stock level strategy, we identified an estimated avoidable overstocking of 173.48 KG across all products during the analysis period. This surplus stock, particularly in high-wastage items, represents an opportunity to reduce wastage and improve inventory efficiency. Implementing the optimized reorder system could significantly mitigate these losses and improve profit margins.

4 Interpretation of Results and Recommendations:

4.1 Increase Sales, Profit, and Revenue:

Table 2 summarizes key business metrics for a vegetable seller over a period of two months. It shows an average daily revenue of ₹3,979.99 and profit of ₹1,033.19, with a recommended target profit margin of around 35%. Positive correlations suggest that increasing profit margin and product variety can improve profits, while weekends consistently show higher demand, and tomatoes, beans, and brinjals are the top revenue drivers.

	Metric	Value
0	Average Daily Revenue	₹3,979.99
1	Average Daily Profit	₹1,033.19
2	Recommended Target Profit Margin	~35%
3	Profit Margin vs Profit Correlation	+0.147 (weak positive)
4	Seasonality Insight	Weekends show consistently higher demand
5	Top Revenue-Driving Products	Tomato, Beans, Brinjal

Table 2. Analysis of Problem statement 1

Recommendations:

1. **Target a 35% Profit Margin**
   Maintain ~35% margin to consistently achieve daily profits of ₹1350–₹1450, leading to monthly income of ₹38,000–₹40,000.
2. **Expand Product Variety**
   Offering 15+ types daily increases profitability. Include high-margin specialty vegetables like red/yellow capsicum, baby corn, and mushroom.
3. **Use Predictive Demand Forecasting**
   Implement Time-series forecasting weekly to align stock with expected demand and reduce missed opportunities.
4. **Introduce Smart Pricing Adjustments**
   Dynamically tweak prices of top-moving vegetables based on forecasted demand or seasonality to boost margins.
5. **Start Street Hawking techniques**
   Street hawking reduces operational costs like rent and utilities, increasing overall profit margins. It also allows flexible location-based selling, boosting sales in high-demand areas.

4.2 Minimize Stock Wastage:

- **Top Perishable Items**: Gooseberry, Beans, Mushroom — over 40% average wastage
- **Correlation between Overstock & Wastage**: +0.63 (moderately strong)

Table 3 highlights the impact of reducing overstock and wastage, showing that 27.38 KG of overstock could be avoided, leading to 21.90 KG less wastage. This reduction results in ₹657 saved in wastage costs and a potential profit gain of ₹328, with high forecasted revenue accuracy.

	Metric	Value
0	Avoidable Overstock (KG)	173.48 KG
1	Estimated Avoidable Wastage	138.78 KG
2	Wastage Cost Saved	₹4377.70
3	Potential Profit Gain	₹1445.17
4	Correlation: Stock vs Wastage	+0.63

Table 3. Analysis of Problem statement 2

Recommendations:

1. **Avoid Overstocking High-Risk Items**

   Products like carrot and capsicum should be stocked 20% - 30% less than others due to spoilage risks.

2. **Visual Alerts for Overstocking**

   Set alerts if stock purchased exceeds optimal threshold - especially for high-wastage SKUs.

3. **Utilize ABC Analysis for Stock Prioritization**

   Focus inventory control efforts on ‘A’ class products that contribute to 80% of revenue.

4. **Bundle or Discount Surplus Perishables**

   Use short-term pricing offers to sell items approaching spoilage and reduce inventory loss.