INDIAN INSTITUTE OF TECHNOLOGY, MADRAS

BUSINESS DATA MANAGEMENT CAPSTONE PROJECT

MID TERM SUBMISSION

ON

"Promotional offer analysis and stock forecasting at a Supermarket"

Submitted by

YUKTI SETHI

Date of mid-term report submission: 09-05-2023

TABLE OF CONTENT

TABLE OF FIGURES

1. Executive Summary

This mid-term report describes the business data management capstone project for ####Super Store, a grocery store owned by a private company in Okhla, New Delhi, and aims at solving 2 business problems faced by the store, namely, the inability in tracking the effectiveness of discounts on sales and poor stock management due to insufficient planning of sales. This report discusses the dataset obtained from ####Super Store along with the significance of every attribute to make meaning out of the analysis being done. It describes the process of data cleaning and data exploration which includes imputing the missing values, and handling incorrect data cells as well as outliers. Descriptive statistics of the data are analysed which give an insight into the variability and distribution of features. The analysis is mainly focused on a weekly, weekday-wise, and daily basis. It has been observed that putting more discounts on low-selling products on weekends and being up-to-date with the stock levels during the start and end of the month will help the business in increasing revenue. A deep SKU-wise analysis is in the final submission scope of this report. Both Machine Learning and MS Excel are used.

2. Problem Statement

1. ####Super Store is struggling to track the effectiveness of its promotional offers and discounts, resulting in a decrease in profit margins as well as customer loyalty.
2. ####Store is facing problems in the timely procurement of stock for the customer due to poor stock management, thereby leading to a loss of customers.

Objectives:

- To improve the effectiveness of its promotional offers and discounts in sales.
- To forecast the stock based on the sales data.

3. Proof of originality

1. Primary data (raw) of ####Super Store for April 2022: [redacted]
2. Customer Bills: [redacted]. Photographs of store and me: [redacted]
3. Letter from the organisation: [redacted]
4. Video of interaction with the employee link: [redacted]
5. Documentation of meetings with proof link: [redacted]

4. Metadata and descriptive statistics

Metadata

The manager of ####Super Store provided sales data for April 2022. The data was collected from the Point of Sale (POS) systems of the superstore.

- File name: “April2022.xlsx”.
- Data Source: ####Super Store POS system
- File format: Excel (.xlsx)
- Name of the sheet: DATA
- Date created: 29-04-2023
- Date last modified: 29-04-2023
- Language: English
- Time frame: 01-04-2022 to 30-04-2022
- Description: One-month sales data to solve business problems at ####Super Store.
- Size of the dataset: 53848 rows × 13 columns
- The rows describe each bill generated on a particular date for buying a specified quantity of a particular stock-keeping unit.

Attributes	Data Type	Unit	Significance
Date	Date/time	Datetime(“dd-mm-yyyy”)	Date when the transaction was made
Bill number	Text	Alpha-numeric	Unique identifier for each transaction
Particulars	Text	Alpha-numeric	Whether the purchase was made by local customers or any other Business organisation
Item	Text	Alpha-numeric	Name of the item purchased (Stock Keeping Unit)
MRP	Number	Currency (Rs.)	Maximum retail price of the item
Discount	Number	Currency (Rs.)	Discount offered on the item
Sale price	Number	Currency (Rs.)	Price at which the item was sold (GST excluded)
Quantity	Number	Number	Number of units purchased
GST amount	Number	Currency (Rs.)	The amount of Goods and Services Tax (GST) charged on each transaction
GST%	Number	Percentage	The percentage of GST charged on each transaction
Net amount	Number	Currency (Rs.)	Total amount paid by the customer
Buying price	Number	Currency (Rs.)	Cost of purchasing the item from the supplier

- Assumptions:
    1. It is assumed that the item names and descriptions in the data are consistent with the store's actual inventory.
- Limitations:
    1. The data is limited to transactions processed through the POS system and does not include sales made through other channels (e.g., online, phone orders).

Descriptive Statistics

On using Pandas, NumPy, matplotlib and seaborn, the number of unique values in each feature and their datatypes are carefully observed. It is found that:

Date	30
Vch/Bill No	7803
Particulars	4
Item Details	5744
MRP	437
Discount	195
Sale Price	1543
Qty.	114
GST Amount	1368
GST %	4
Net Amount	608
Buying Price	2845
VAL	849

Note: Each row in the raw dataset contains a given quantity of SKU bought by the customer at a given date as part of a particular transaction (denoted by bill number), along with other attributes. Rows with the same bill number and date represent the same transaction, but the bill number does not uniquely identify each row.

Figure 1 Descriptive Statistics: Number of unique values in each feature of the dataset.

Item details: Total number of unique items in the store is 5744. These items are, in fact, stock-keeping units and may differ by the amount of quantity present in them as well. For example, 'FUNFOOD MAYO VEG 250G' and 'FUNFOOD MAYO VEG 400G' are 2 different SKUs and are thus counted uniquely.

Bill Number: 7803 unique bills are billed in April 2022 for the customers visiting the superstore in the same period. The bill number is just a unique identifier does not provide any essential information about the customer.

Particulars: It includes 4 values: 'Cash' refers to all the payments being made by the local customers visiting the store either in cash or online but are essentially payments made at the store in person and not a completely phone/online order so to say. 'Human Welfare Foundation', 'Human Welfare Trust', and 'K S Ultimate (17-18)' NGOs that give large amounts of orders to the store.

Date: It is in Datetime format and has 30 unique values. These 30 values account for 30 days of April 2022 sales.

The Discount column has values in rupees in purchases where the customer has got a discount. Otherwise, the cell is just left blank. This accounts for blank values in the discount attribute as well. There is a mean discount of Rs 8 in the dataset.

The 'Quantity' attribute can also be a float because the store deals with products like pulses and flour that are measured in grams and kilograms. The maximum number of a particular product bought on a given day is equal to 2000 (due to bulk orders) while the interquartile

range happens to be constant at 1 since customers usually buy 1 unit of a particular product at a time.

MRP		*Sale Price*		*Qty.*		*Net amount*		*Buying Price*		*Val*	
Mean	111.7755	Mean	93.43809	Mean	1.951467	Mean	103.2044	Mean	81.14549	Mean	177.6805
Standard Error	0.647545	Standard Error	0.505755	Standard Error	0.090771	Standard Error	0.554084	Standard Error	0.455955	Standard Error	12.59644
Median	70	Median	61.9	Median	1	Median	65	Median	49.58	Median	80
Mode	10	Mode	17.86	Mode	1	Mode	30	Mode	7.7	Mode	20
Standard Deviation	150.2597	Standard Deviation	117.3581	Standard Deviation	21.06309	Standard Deviation	128.5726	Standard Deviation	105.8021	Standard Deviation	2922.943
Sample Variance	22577.98	Sample Variance	13772.93	Sample Variance	443.6539	Sample Variance	16530.9	Sample Variance	11194.09	Sample Variance	8543596
Kurtosis	269.2809	Kurtosis	74.96397	Kurtosis	4277.263	Kurtosis	66.0352	Kurtosis	74.15007	Kurtosis	8263.125
Skewness	9.61107	Skewness	5.861589	Skewness	59.85246	Skewness	5.536638	Skewness	5.810348	Skewness	84.18761
Range	7193	Range	3806.89	Range	1999.95	Range	3997	Range	3474.11	Range	329998
Minimum	2	Minimum	1.69	Minimum	0.05	Minimum	2	Minimum	0.01	Minimum	2
Maximum	7195	Maximum	3808.58	Maximum	2000	Maximum	3999	Maximum	3474.12	Maximum	330000
Sum	6018550	Sum	5031174	Sum	105076.8	Sum	5557040	Sum	4369279	Sum	9567208
Count	53845	Count	53845	Count	53845	Count	53845	Count	53845	Count	53845

Figure 2 Descriptive statistics of numerical features.

MRP: It is the maximum retail price of the product and a discount is applied on the MRP of the product. The MRP includes the GST prices as well. The maximum MRP among all the products is Rs. 7195, while the average is Rs. 70.

Sale price: It is the selling price of the product without including GST but including the discount, if any. The maximum sale price of the product is Rs. 3808 while the average is Rs. 61.

Net amount: It is the amount at which one unit of that item was bought by the customer in a particular transaction. It includes any discount that the item is associated with. **Net amount = MRP – Discount. Net Amount= Sale price + GST amount**. The average sales net amount gained from each product is Rs. 65 which is lesser than the average MRP (Rs. 70). This is due to discounts.

Buying Price: It is the price at which the store purchased a one-unit quantity of that product from the vendor. The average buying price of each product is Rs. 49.

Val = Net amount × Quantity. Value is thus the total revenue generated from that particular transaction of a particular item for the quantity purchased by the customer. The average revenue generated from each transaction of a particular quantity of a given product is Rs. 80. However, the maximum revenue touches Rs. 330000.

Figure 3 shows the distribution of quantity bought per item in one month and is mostly concentrated near 1 along with many outliers which are high-volume items. Similarly, the distribution of the volume of products sold by date has an outlier that touches the quantity of

about 23000 products sold in one day. However, after removing outliers, the average quantity sold per day is about 2300 (Figure 4).

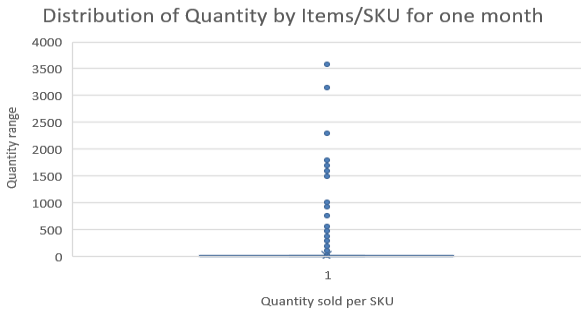

Figure 3 Box plot showing distribution of quantity sold by items for one month.

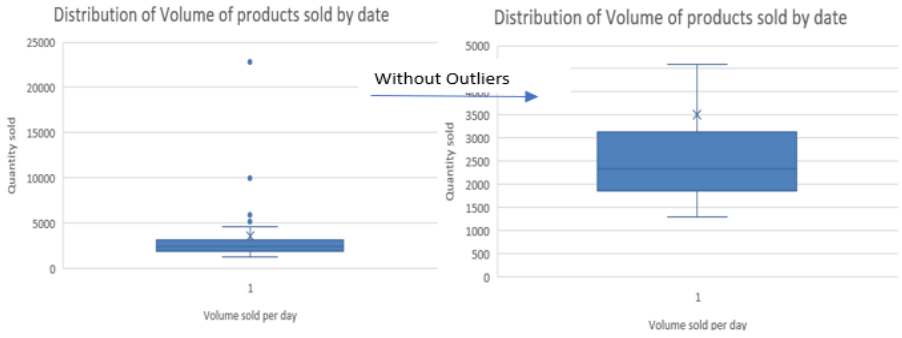

Figure 4 Box plot showing distribution of quantity of products sold by date with and without outliers.

The distribution of revenue by date shows that there was a huge spike in revenue and Rs. 25,00,000 amount of revenue was made in one single day. Similarly, the distribution of revenue by Items shows that certain high-revenue generating products are a cause of outliers in the distribution (Figure 5).

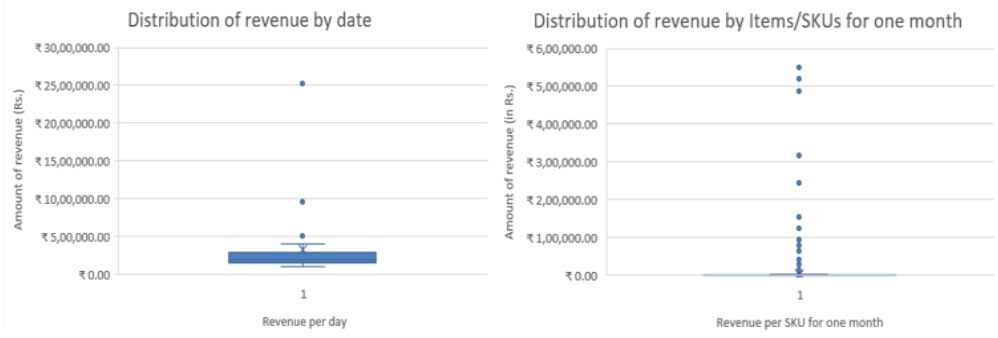

Figure 5 Box plots showing distribution of revenue by date and by items for one month.

The correlation of the features: dark colours give a strong negative correlation while the light colours give a strong positive correlation. It is observed that the sale price has a high linear correlation of 0.97 with MRP. This means that as the sale price increases, the MRP of the product also increases. Qty (quantity of products sold) also has a positive correlation of 0.81 with VAL (revenue generated) which is obvious. Notice that the Net amount has a slight

negative correlation of -0.0088 with Quantity. This is due to the increased discounts on large-volume items. However, it is a negligible value. (Figure 6)

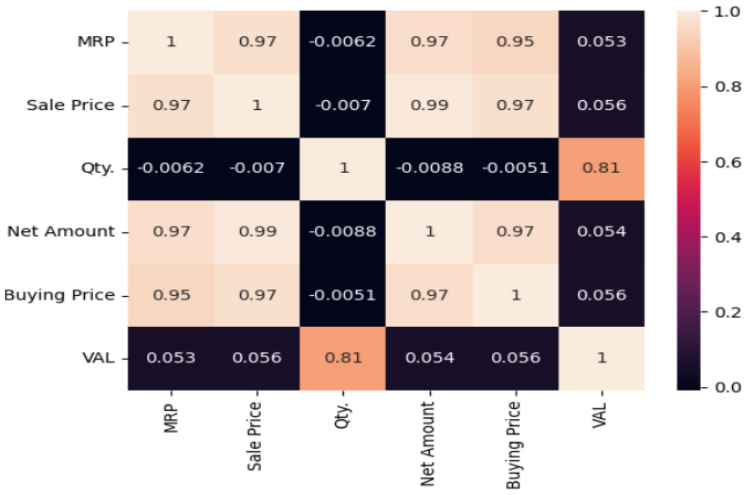

Figure 6 Correlation heatmap between numerical features in the dataset.

5. Detailed explanation of analysis/method

1. Data cleaning

In this step, the data is cleaned appropriately by searching for anomalies and outliers. Pandas, NumPy, matplotlib, and seaborn libraries are also used for quick analysis and data cleaning. The following steps are taken:

a. Removal of missing values

It is found that the sale price and cost price essentially contain certain missing values due to errors in data capture in the POS systems of the superstore. By using the sort and filter tool of MS Excel, 2 rows containing missing values in the 'Sale Price' column are dropped because they are too less to affect the analysis. However, there are 286 rows containing missing values in the 'Buying Price' which will be a loss of valuable data if dropped Thus, Regression analysis is used to predict the buying price of the products using the selling price. for which it is found to be missing. This is done using 2 tools: **Analysis Toolpak add-in in MS Excel**: It gives regression statistics by analysing the data and thus gives the coefficients of the linear relationship between the Net Amount and the Buying Price. Taking 'Buying Price' as the dependent variable and 'Net Amount' as the dependent variable, an equation of the form Buying Price= Net Amount * coefficient + intercept is obtained. **Pandas, NumPy, and sklearn**: To check the accuracy of the linear model for missing value imputation, the same model is implemented using the Linear Regression model in sklearn that gives about 0.93 r2

score on train data and 0.92 r2 score on test data. It, thus, validates the estimates made by the linear equation to be, indeed good approximates. The predicted values are incorporated in place of the blank fields in the 'Buying Price' feature.

b. Handling anomalies:

It is found that quantity has certain negative values. On consulting the owner regarding such an error, it is found that this is due to a certain error in the POS system which reflects that the products are out of stock but the product, is in fact, right there in the customer's hands. Thus, the system is made to bypass the error to make a transaction which is why certain fields in the quantity column are negative. These are handled by converting them to positive values by applying the absolute function on the column.

2. **Weekly sales Analysis**

Using the cleaned data, revenue generated weekly is analysed and the increase or decrease in total revenue across weeks is studied. This analysis will help in the allocation of discounts at the right time when the sales are low and thus will attract more customers. This peak activity time will also help in selling those products which are not selling much, by putting them on discount, thereby addressing the problem of promotional offer analysis. The variation will also help in prioritising the stock level of products during this peak time of sales, thereby addressing the problem of stock forecasting.

3. **Daily volume trend**

The volume of sales on each day is studied using the moving average that takes the average of the quantity sold in the past 4 days in order to smooth out the fluctuations in sales data and identify spikes or dips in sales. This helps in prioritising the availability of items based on the daily trend and forecasting stock demand by projecting the trendline in future.

4. **Analysis of revenue-generating sources**

The total revenue generated from various sources is analysed by creating a pivot table with 'Particulars' as the rows and revenue as the value. This is done to study the distribution of revenue across sources and to consider the impact of removing the sources which lead to a sudden spike in revenue at a given day. In other words, the impact of removing the anomalies on sales is studied before removing it from the analysis.

5. **Predictive analysis of profit and revenue on weekdays with high local customer sales to allocate discounts on weekdays.**

The dataset is grouped on days of the week using a pivot table and a clustered column chart is generated to analyse the amount of revenue generated on those days along with profit. This will give the incentive to give weekday-specific discounts to increase the engagement of the store and thereby increase the revenue. This analysis is done after removing the outliers associated with bulk orders from non-government organisations to take the weekday-wise variation of the general customer base into account. The change in profit is observed by using the formula Profit = (Selling Price- Buying Price)/Buying Price.

6. Results and Findings (Excel analysis: [redacted])

1. **Regression analysis while missing value imputation**

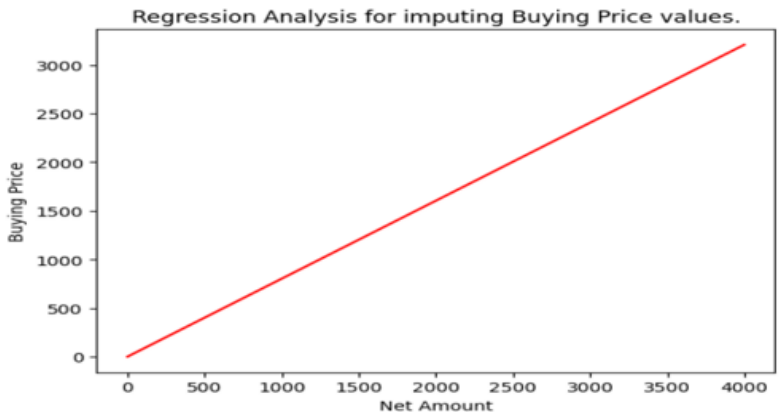

Figure 7 Linear regression model for Buying Price prediction using Net amount.

A relationship "**Buying Price = 0.80* Net Amount – 1.64**" is found while using regression analysis in both Excel and Machine Learning. The relationship is 92% accurate which is a good confidence interval. The mean squared error of training and test data match thereby indicating that the model is correctly fit. Click here: regression analysis.

2. **Weekly Revenue analysis**:

It is found that the total weekly revenue generated by the store is high during the initial week of the month after which it decreased by 59% during the second week. However, the revenue spiked to about Rs. 36,00,000 during the third week (241.1% increase from the previous week). On consulting the manager about this sudden spike, it is found that the invoices of the orders from NGOs were booked on 18th April 2022 as the organisations

generally order in bulk during Ramadan. After this week, the line chart shows decreased revenue generation, which, essentially is the actual revenue distribution of the store with only the local customer base in consideration. (Figure 8).

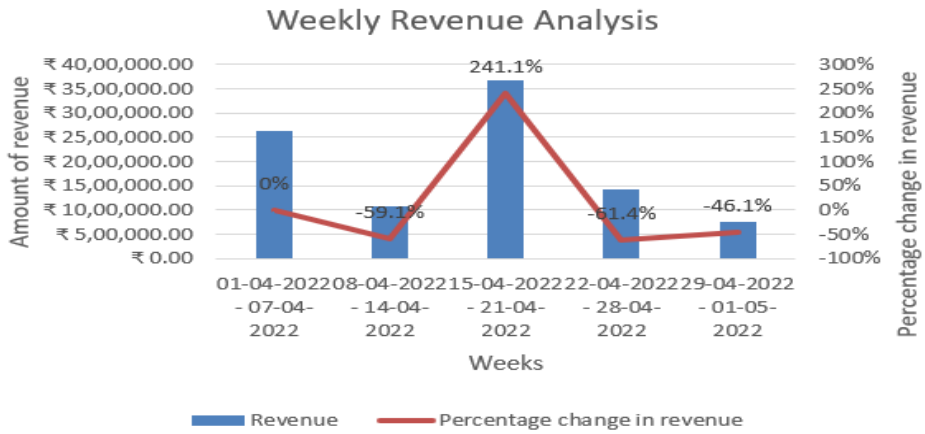

Figure 8 Weekly revenue analysis with % change in revenue.

3. **Daily Volume trend**

After removing the anomaly, the trend in local customer sales and moving average of quantity of products sold in the past 4 days shows that there is a significant increase in sales volume at the start and end of the month. The stock availability should be prioritised at this time or the store should at least have a huge safety stock at hand on these days which see a high influx of customers and increasing sales. (Figure 9)

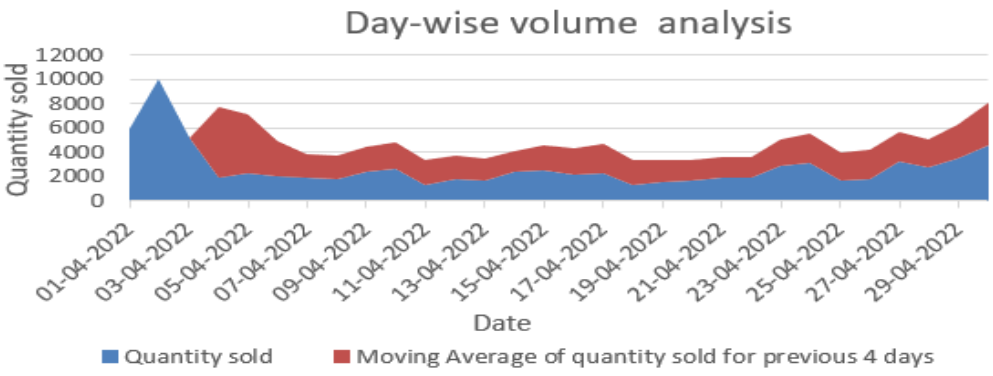

Figure 9 Day-wise volume analysis for one month.

4. **Revenue generating sources**

The below graph shows about 27% of the revenue of the store came from supplying groceries to non-governmental organisations in April 2022. However, the actual distribution of revenue from local customers' accounts for a total of Rs. 70,00,000 per

month approximately, if the seasonal fluctuation due to Ramadan did not have a significant impact on the local customer expense base of the store, which would otherwise make the total revenue lower than what has been observed in April 2022. (Figure 10)

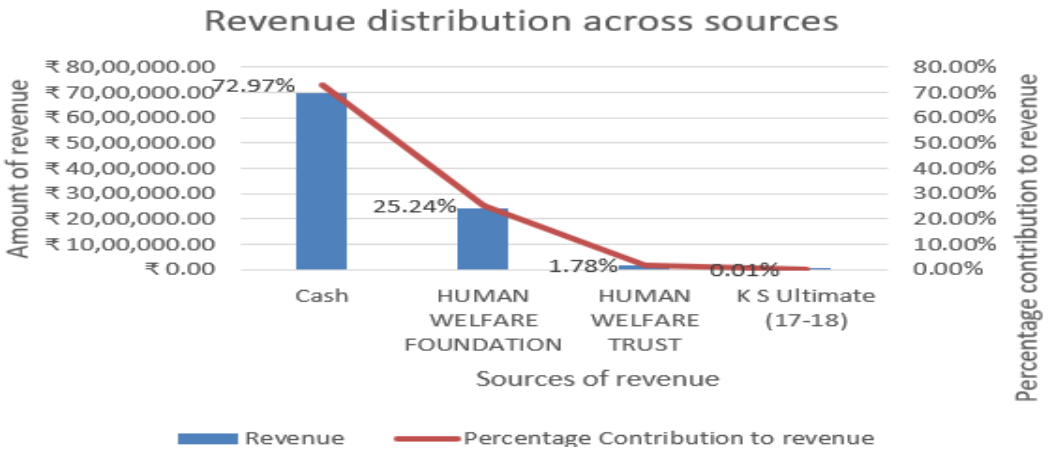

Figure 10 Revenue distribution across sources and their contribution to total revenue of the store for one month.

5. **<u>Predictive analysis on weekdays with high local customer sales</u>**

The local customer sales of the store are significantly higher on Saturday followed by Friday and Sunday with a profit of 25%. In this case, a discount on the high-selling products (which will be analysed later in the final report) will be decreased specifically on Saturday to generate more profit. Stock of high-revenue-generating products should also be up-to-date till the weekend for greater revenue. (Figure 11)

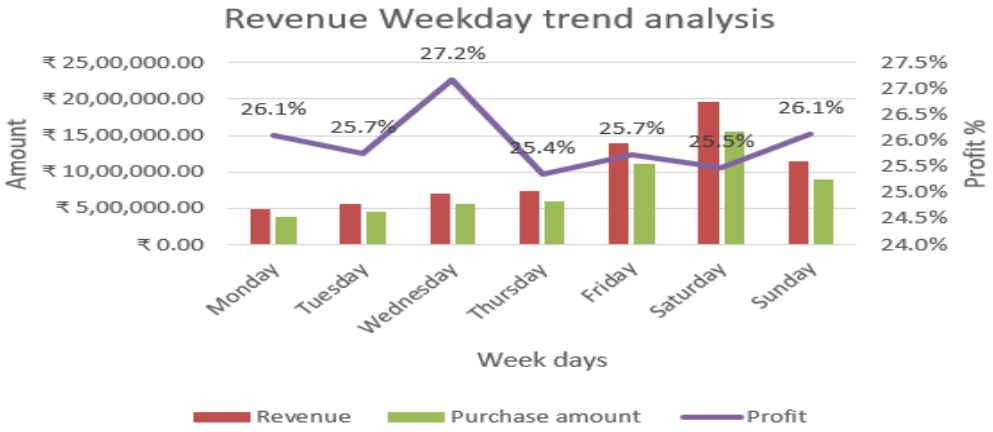

Figure 11 Revenue Weekday Trend analysis along with Profit%.

<u>Way Forward:</u> SKU categorisation for visualisation, SKU with discount vs without discount comparison, High-volume categories analysis to forecast stock, levying discounts on low-profit & low-revenue-generating products, Reorder point analysis using sales data.