IIT Madras
BSc Degree

Diploma in Data Science

Business Data Management

(BSCMS2001)

Capstone Project Final Submission

Estimating Uncertainty in Supermarket Sales During Festive and Non-festive Seasons

Submitted by

Durga Shree N

(21f2000610)

Estimating Uncertainty in Supermarket Sales During Festive and Non-festive Seasons

1. Executive Summary

The purpose of this project is to help a supermarket deal with a few of its challenges by providing useful insights from everyday sales and purchase data of two months, October and November. This organization, a relatively new one in the locality, is looking for strategies to gain popularity and trust from the public in par with other old stores. This project aims to summarize the data through visualizations and point out a chuck of products that require alteration in service strategies using pivot tables and defining function logics on dataframes. It is also expected to derive insights on the uniformity of the quantity of products sold and quantify the deviations through standard deviation. Determining the elasticity of quantity of products sold during festive and non-festive seasons supports the shopkeepers with deciding on the right quantity of stocking and the timing for stocking. It is understood from the interaction with the shopkeeper that re-stocking has been done on rough estimations and informal sale pattern observations which has led to understocking and overstocking at times. It's the peak time for the store to not lose customers and start gaining better profits as well and make the right choice about dealers for supplies. Hence, quantifiable outcomes for amending the service strategy and maintenance are expected to be developed from this work.

2. Detailed Explanation of Analysis Process/Method

There are 61 days in the dataset (Number of days in October + November). These days are split into 9 weeks with only 5 days in the last week for the purpose of analysis.

Week 4 occurs for the dates October 22nd to October 28th and Diwali 2022 happens to be in the same week. Hence, this week is a point of interest for all comparisons to identify changes in sales based on festive and non-festive seasons.

Datasets collected:

- Purchase data for the month of October 2022
- Sales data for the month of October 2022
- Purchase data for the month of November 2022
- Sales data for the month of November 2022

Sales data and purchase data for all months follow the same structure.

2.1Extracting the chunk of products that show huge deviation in the sales pattern and predicting re-stocking pattern

Sales data description considered for analysis:

Duration: 2 months (Each file contains sales data (in billing format) of one month).

Data is collected for the months of October and November 2022 given the presence of Diwali festival and its influence in grocery purchase. The data from both months are compiled together into one file including columns for the number of week and day of the week and month for further analysis.

Attributes in the dataset:

Each row corresponds to a product and a bill with multiple products contains multiple rows for different products with the same bill number.

1. **Bill No.:** Generated automatically for every bill
2. **Date:** Date of product purchase
3. **Day:** Day of product purchase
4. **Week:** The 2 months considered are divided into 9 weeks and this attribute include the week number
5. **Month:** Month of product purchase
6. **Product name:** Name of the product with brand and quantity
7. **Quantity sold:** Number of packets
8. **Price:** Price of the product for 1 quantity
9. **Tax%:** Percentage of tax applicable for each product
10. **Rate:** Price of the product for 1 quantity with tax

11. **Total:** Total rate based on quantity

Purchase data description considered for analysis

Duration: 2 months (Each file contains sales data (in billing format) of one month).

Data is collected for the months of October and November 2022 as the shopkeepers would expect sales variation for Diwali festival and increase stocking of certain products. The period also included non-festive season after Diwali. The data from both months are compiled together into one file including columns for the number of week and day of the week and month for further analysis.

Attributes in the dataset:

Each row corresponds to a product in the dataset. Based on the number of different products purchased from a dealer, so many rows will be corresponding to one purchase from a dealer's invoice number.

1. **Date:** Date of order placed
2. **Month:** Month of order placed
3. **Day:** Day of order placed
4. **Week:** The 2 months considered are divided into 9 weeks and this attribute includes the week number
5. **Invoice number:** Invoice number from the dealer
6. **Product:** Product name
7. **Quantity:** Number of entities purchased
8. **Rate:** Price of the product
9. **Net value:** Total price for the number of the quantity purchased
10. **Tax amount:** Tax amount based on tax percentage for the product
11. **Total amount:** Net value + Tax amount
12. **Source/dealer:** The organization name of the product supplier

Number of week days in each month

Day | Month	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
October	5	4	4	4	4	5	5
November	4	5	5	4	4	4	4

This analysis is proposed to be carried out in two ways:

1. Consider sales in each day for 2 months
2. Group the days into weeks and consider deviation between weeks

Step 1: Create a pivot table from data with the product name as rows and week number or date as column and quantity as value

Step 2: Open this table as a Python dataframe

Step 3: Apply box-plot analysis for each product's quantity and extract the product as outlier if the sales in week 4 is classified as outlier

Step 4: Identify the standard deviation of each product for each week

Step 5: Apply box-plot analysis based on standard deviation

Step 6: Extract outliers

Step 7: Perform Steps 4-6 by omitting the festive week (Week 4).

- It can be inferred from the above procedure that the products that appear as outliers in both the festive and non-festive seasons do not generally have a stable sales pattern.
- The ones obtained as outliers only in the festive season can be further analysed to quantify the deviation in quantity purchased as those are preferred products of the festival
- The ones obtained as outliers in the non-festive season can be considered to be those which are generally not purchased for festivals due to ongoing rituals and practices.
- The standard deviation of sales obtained can also be used to quantify how early an order should be placed for the product.

Box-plot analysis

A box-plot analysis is drawn by dividing the dataset of sales quantities/ its standard deviation into quartiles. The set of quantities purchased is divided into 4 quartiles with q1 denoting the first 25%, q2 (median) denoting the 50% position and q3 denoting the 75% position. The difference between q1 and q3 is called the interquartile range (IQR).

Lower bound = q1 – 1.5*IQR

Upper bound = q3 + 1.5*IQR

All the values below the lower bound and above the upper bound are classified as outliers.

2.2 Identifying fast-moving quantity/size of packets of a product

Customers expect preferred sizes of products that contribute high value in their shopping list. Optimizing/ identifying fast-moving size quantitatively for all products would be tedious for a store selling more than 2500 products. Hence, let's consider the 20% maximum value products.

Step 1: Identify 20% of maximum value products from sales dataset.

Step 2: Application of RegEx string matching to identify those products which are available in more than one packet size (This is required as the dataset contains product with different product packet sizes uniquely).

Step 3: Present the percentage of variation between different sizes to the shopkeeper to make informed decisions in the future.

Step 4: If the shopkeeper specifies a certain threshold to consider for packet size optimization, the required products are to be extracted by defining a Python function from the dataframe.

2.3 Python to facilitate analysis

- The pandas dataframe is put to use to handle the data (eg. Traversing).
- Performing outlier analysis of on every product for its everyday and weekly sales requires box-plot analysis. It has been implemented using Python.
- Computing standard deviation in sales of products within every week is again automated using python code. The higher the standard deviation in sales, the more is the unpredictable demand for a product during a peak season.

2.4 Pivot tables to facilitate analysis

The aggregation of data based on constraints such as weeks, months, days of the week is done efficiently using pivot tables before using python for further automated analysis.

2.5 Limitation of the dataset

The dataset the organization was able to provide was a billing database. The dataset is pre-processed to provide a structure for analysis. Since the dataset only consists of product names, quantity purchased, price and tax % for each product, it is quite complicated to perform analysis given the presence of 2500 products in the billing dataset.

If there exist categories to items being sold, analysis on aggregation of products from same category would have provided many inferences.

3. Results and Findings with interpretations

3.1 To interpret if there is an influence of Diwali in revenue generated

The revenue generated in the months of October and November differ by an amount > 1 lakh, which can be attributed to the Diwali festive season in the last week of October.

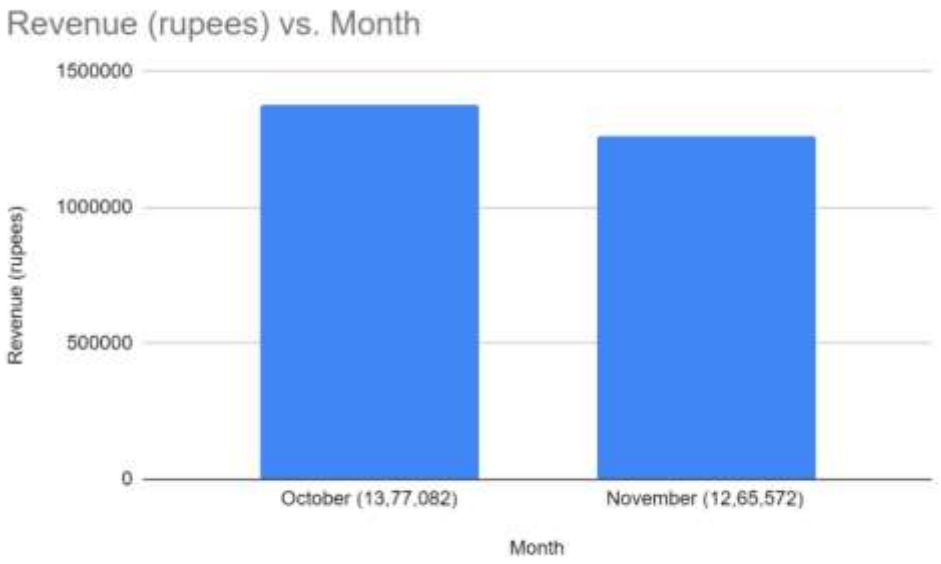

Figure 1: Revenue generated by months with and without festive season

$$Contribution\ from\ the\ festive\ season\ in\ terms\ of\ revenue = \frac{Revenue\ in\ October - Revenue\ in\ November}{Revenue\ in\ November} = 8.8\%$$

The portion attributed to the festive season in fig. 2 also includes the daily basis purchases. Hence, it cannot be interpreted that 3 lakhs+ of revenue is generated from the festive season alone.

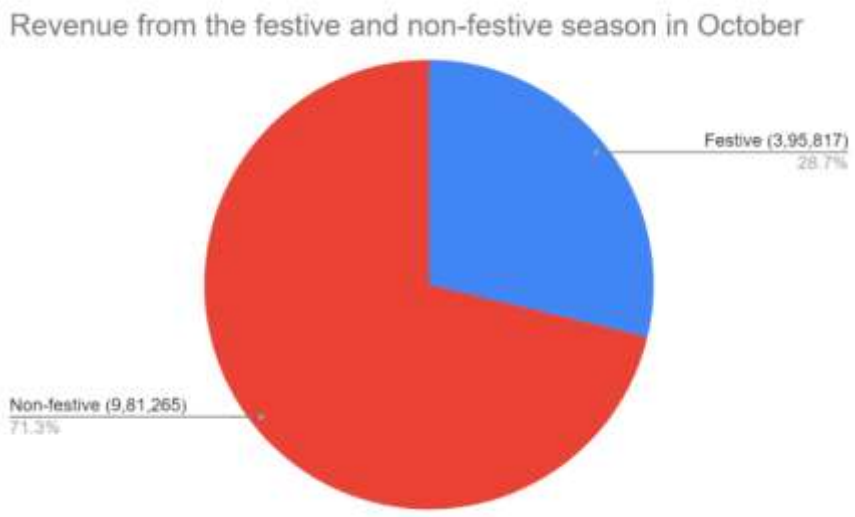

Figure 2: Revenue generated in the festive week vs the other weeks together

It can be observed that there is no huge role of festive season in the revenue distribution for the month of October.

3.2 Identifying a few items that were sold in higher quantities for Diwali

The 2.1 section of detailed analysis process/method aims to interpret the products in huge demand through the concept of box plot analysis. The extracted outliers are those whose sales show huge deviation from normal sales. The outliers are extracted in three ways:

1. Everyday sales quantity of the product
2. Weekly sales quantity of the product
3. Standard deviation in the sales quantity in a week of the product

Since the interest is to concentrate on those products that are in huge demand during festive seasons, the outliers from the festive season alone can be considered. A few products which are in very high demand appear as outliers in all the three outlier extraction mechanisms.

These products include:

- Gram flour

- Deepam oil
- Idhayam oil
- Karpooram (camphor)
- Nattu vellam (special jaggery)
- Vellam (jaggery)
- Soan papadi
- 3 Roses
- Gold winner oil

Figure 3 shows those products that were picked as outliers based on quantity purchased along with the increased contribution to the revenue.

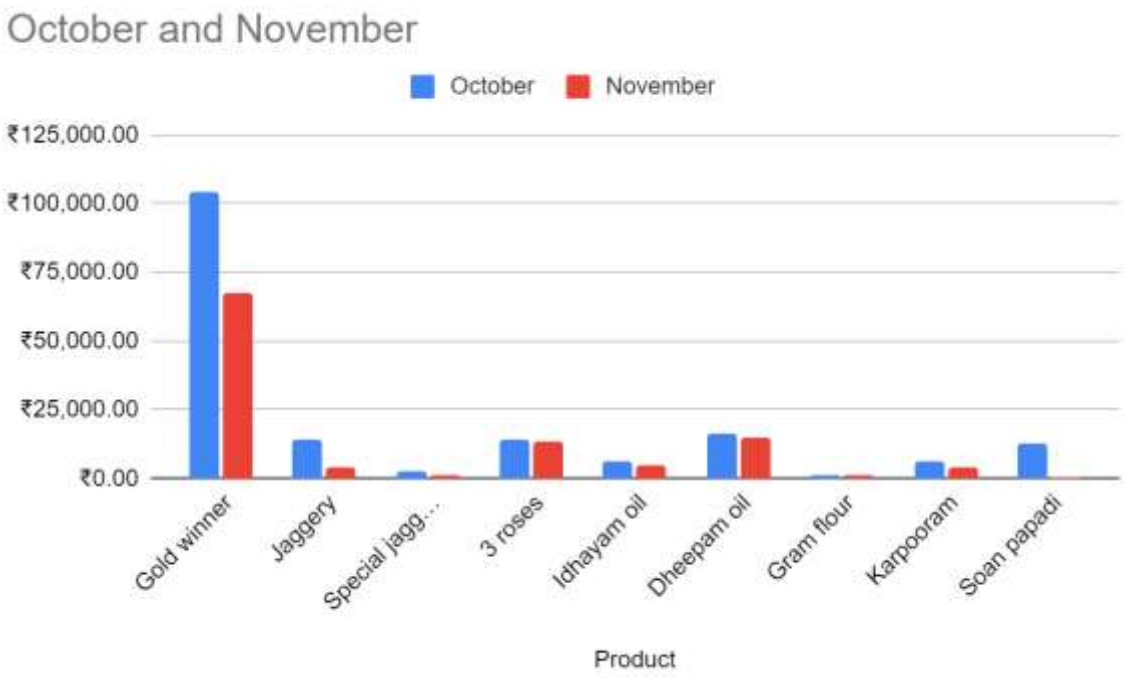

Figure 3: Items identified as outlier for increased sales and their corresponding revenue

3.3 When should these products with huge demand be restocked?

Considering the instance of gold winner oil because of its highest contribution to revenue, it can be observed that the sales is at peaks only on the first two days of festival.

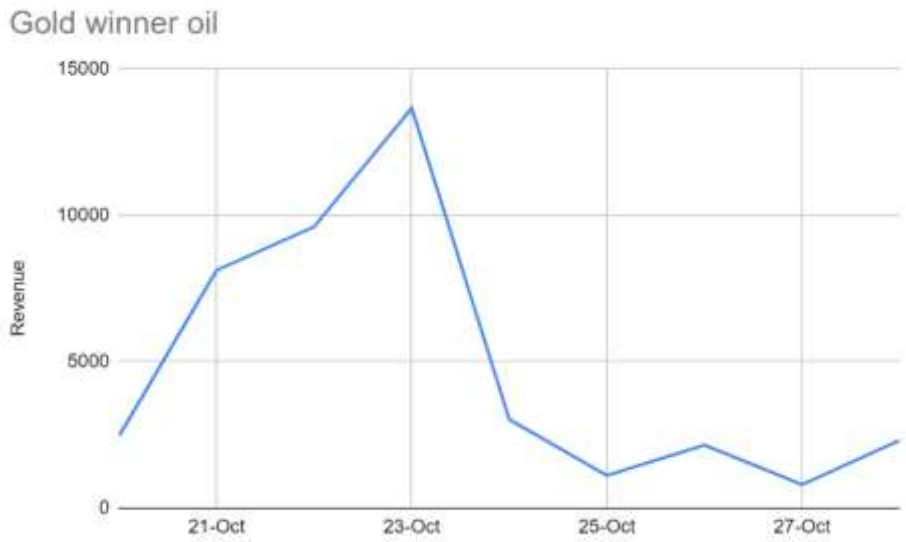

Figure 4: Revenue from sales of gold winner oil in the festive week of Diwali

Similar representations can be made on the other outlier demand products as well. Hence, ensuring stocks for these days is quite important to avoid stockouts during high demand. Products in high demand should be restocked well in advance before the festive season.

3.4 Is the increased sales having a role in the tax% generated?

Figure 5: Products with a tax% and its corresponding revenue generated

It can be inferred that the revenue generated from products with different tax percentage remains the same across both the months. The presence of festive season in October has not affected the tax-based revenue generated.

Those products with 5% tax show a little increased revenue in the month of November. Food products are prone to 5% tax and with increase in purchase quantities, it makes sense for a slightly increased revenue during October.

3.5Fast-moving quantity/size of packets of a product

The shopkeeper cannot concentrate on every item to maximize the capacity of those sizes generally preferred by the customers. Hence, the top 20% revenue generating products can be considered by the owner for his turn-over improvement as well. On evaluating the dataset, it is inferred that those packets in minimal sizes are purchased a lot frequently.

For example, gold winner oil packets of 1 litre are purchased more frequently than 5 litre cans; masala powders of 50 gms are preferred over ones with 100gms or 200 gms, etc.

3.6 Visualization of the sales and purchase patterns in two months

Weekdays vs. Sales for each month

It can be observed that high sales are observed on Sundays. The number of Sundays in the month do not significantly contribute to the sales amount of Sundays. It can be observed that the contribution in overall sales value increases for working days if the number increases in the month (Example: Monday and Wednesday).

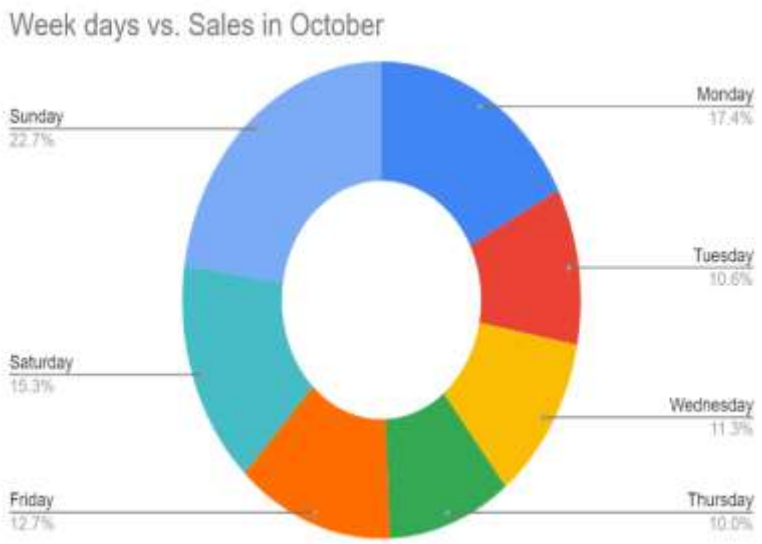

Figure 6: Contribution of days of the week in sales value at the supermarket for Oct

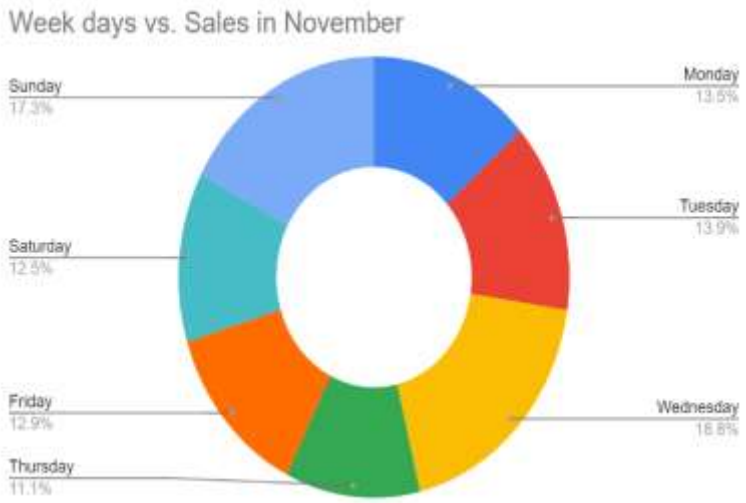

Figure 7: Contribution of days of the week in sales value at the supermarket for Nov

Weekdays vs. Purchase in each Month

Tuesdays specific for Purchase in the town of Denkanikotta. This case was observed a few years back when shopkeepers travel to Bangalore for purchase. This pattern can be clearly observed in the purchase chart of October. Whereas, November shows a significant variation.

The shopkeeper mentioned about the increase in interest of owners to have dealers who supply stocks to the supermarkets directly post-covid. Dealers do not have specific days for supplying goods. Hence, the purchase will be spread out on all the days. The November Purchase chart reflects this.

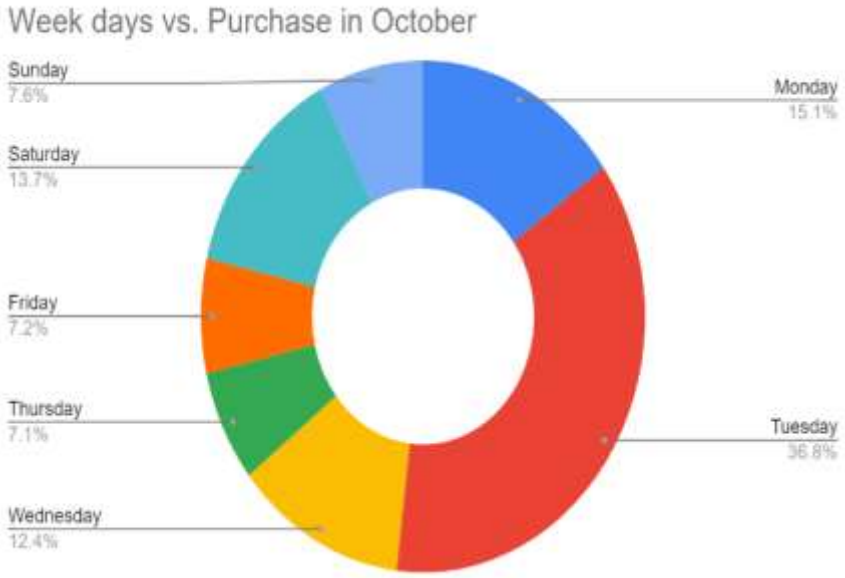

Figure 8: Contribution of days of the week in purchase value at the supermarket for Oct

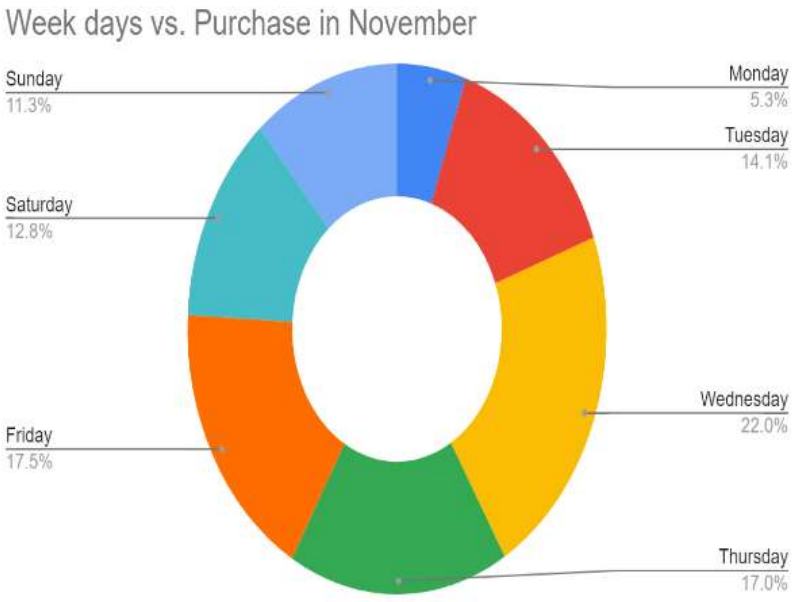

Figure 9: Contribution of days of the week in purchase value at the supermarket for Nov

Sales vs. the week

The 2 months (61 days) are split into 9 weeks. The last week cannot be considered for effective comparison because of the lesser number of days (5 days).

The 4th week (22nd October – 28th October 2022) is Diwali week. As expected, the sales are high in weeks 3 and 4 (15th October to 28th October). People begin their purchases a week ahead generally and this is evidently reflected as increased sales amount. Week 5 sees a slight drop is sales are festive spendings.

Another observable feature of high sales in the first week of October is that people prefer to do re-stocking in the beginning of every month as they receive their salaries. The reason for not observing this in the first weeks of November (week 5 and 6 29th October – 11th November) could be attributed to the over-stocking of goods by people for the festival.

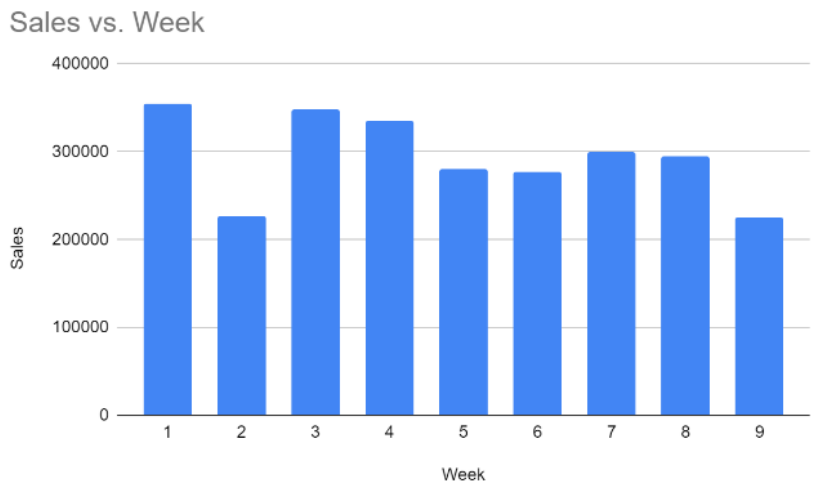

Figure 10: Sales amount and purchase amount for each week

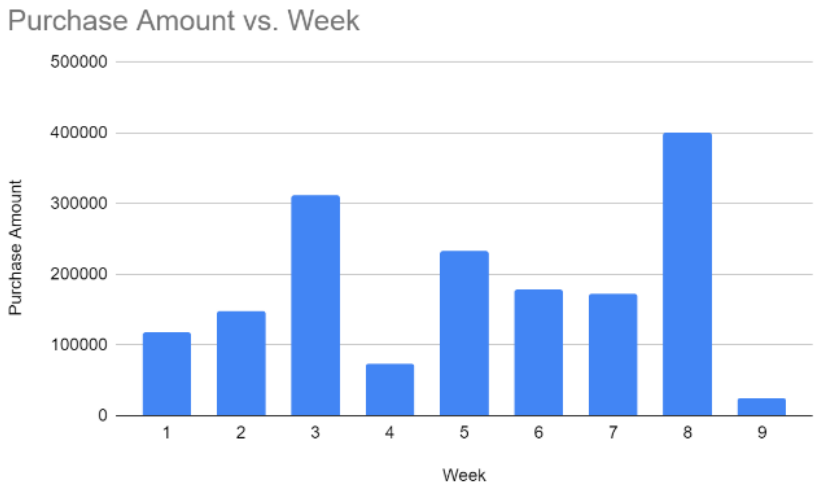

Figure 11: Purchase amount and purchase amount for each week

The purchase value is significantly high in the third week as the supermarket restocks all the necessities for festival sales.

Purchase value and Dealers/Suppliers

The shopkeeper seems to have almost the same value of dealing with the specified dealers/suppliers. Others include the purchase of those products whose dealers are not static and the shopkeepers keep experimenting with various dealers based on price, speed and ease of stocking.

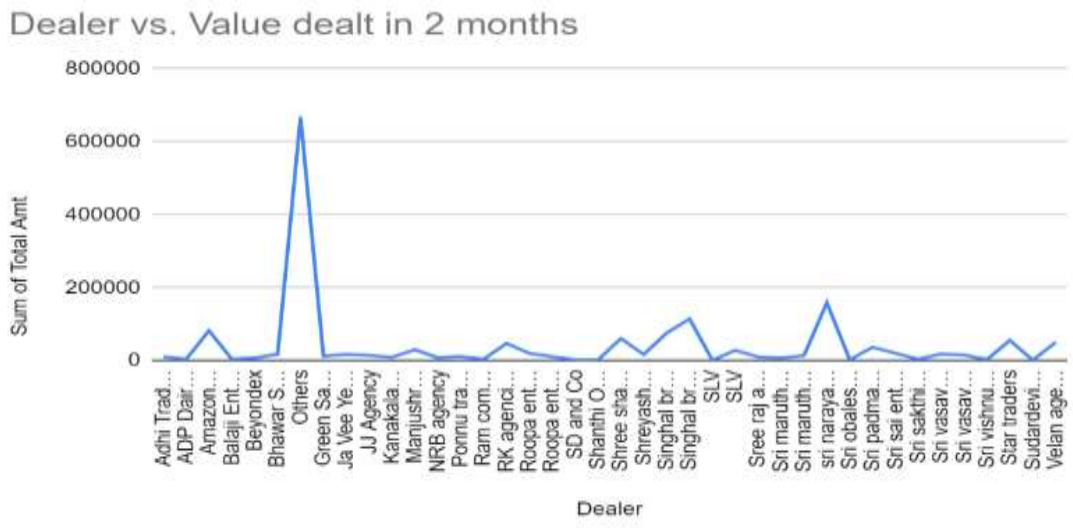

Figure 12: Purchase value of the supermarket with various dealers

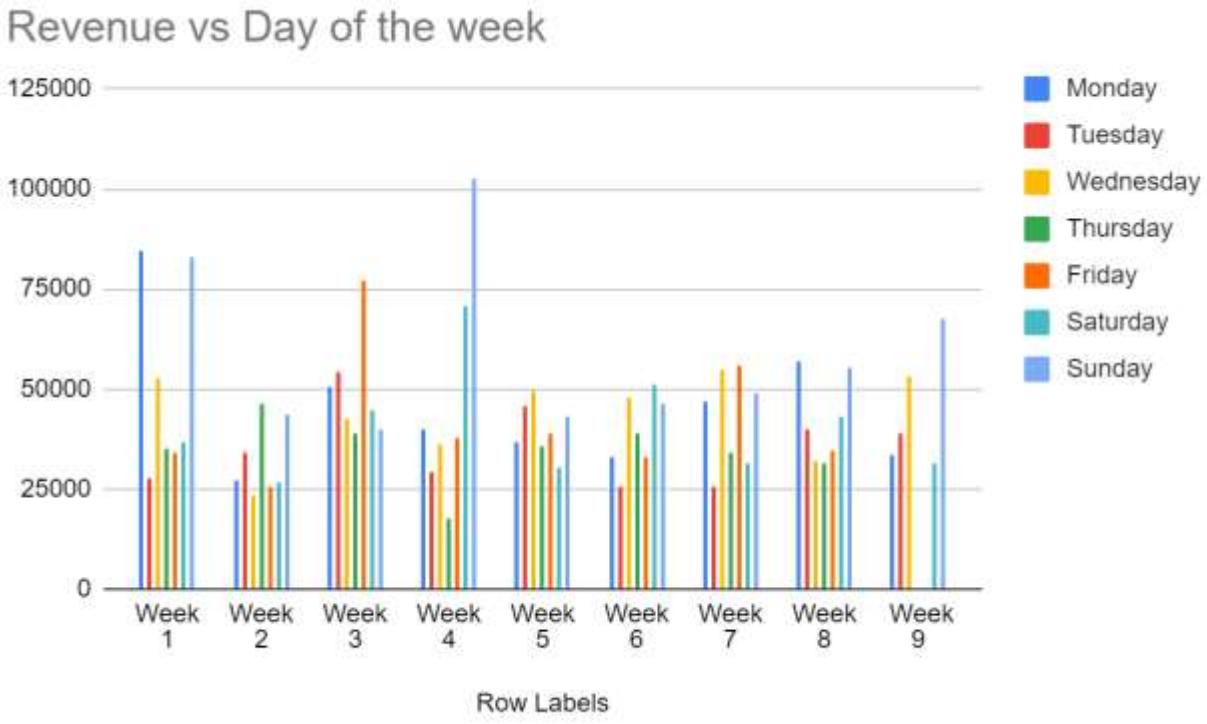

Figure 13: A summary of the sales data analysed based on revenue

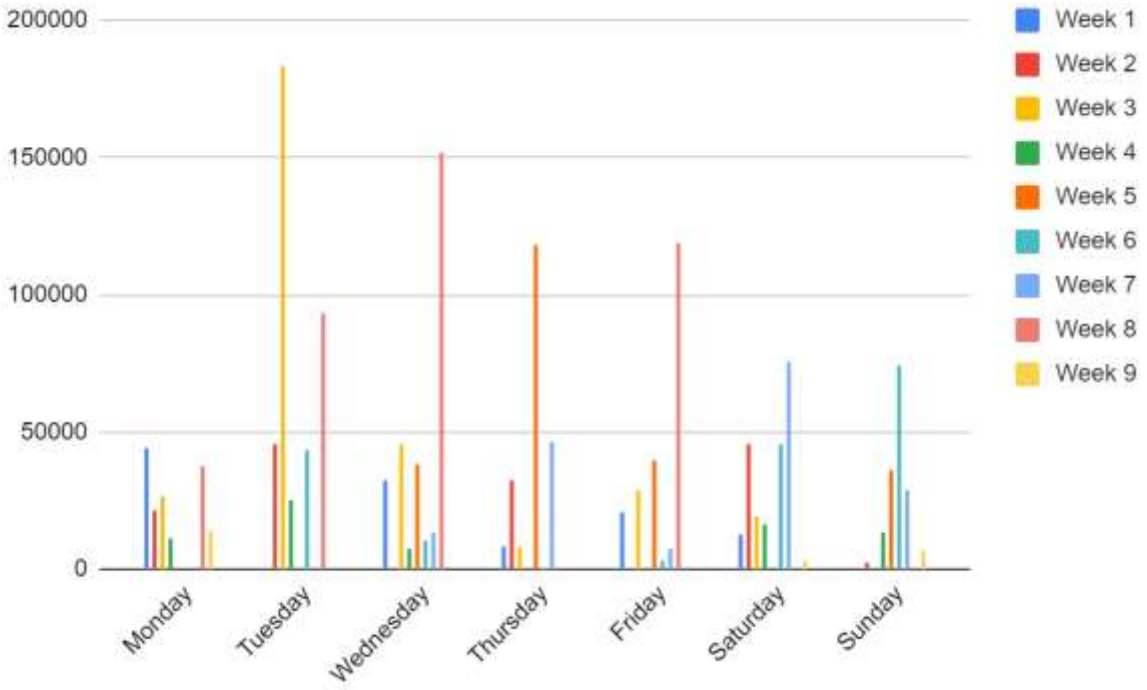

Figure 14: A summary of the purchase data analysed

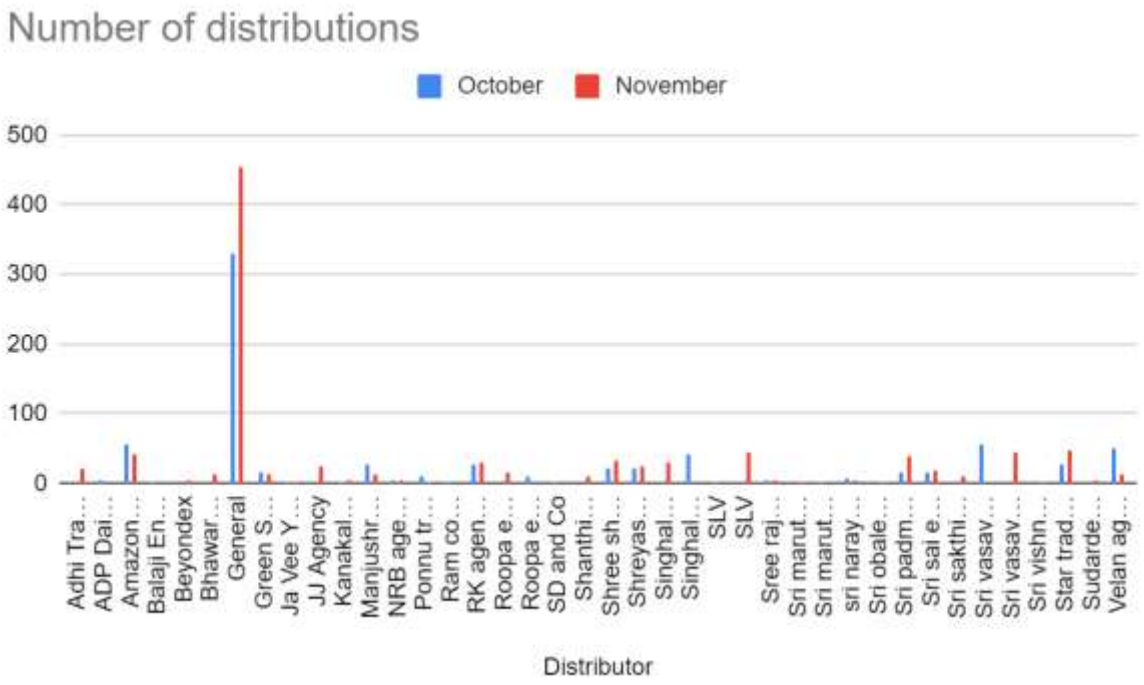

Figure 15: Number of distributions by each distributor in each month

4. Interpretation of Results and Recommendation

Section 3 summarises the finding and also performs interpretations of those results.

The supermarket sells and purchases all types of products. Given the wide range of varieties, it is quite difficult to predict the demand of a product. Groceries have a wide range of purchasing patterns. A distinction of seasons can be used as a mode to analyse the difference in demand of products. On the basis of analysis of the supermarket data during both festive and non-festive seasons, it is recommended

1. To avoid situations of stockouts as initial days of festive season show huge purchases.
2. To prefer low to medium-sized packets for products while customizing the packet size/quantity within the supermarket.
3. To ensure to stock goods by weekends as the sales are slightly higher.
4. Festive weeks generally face huge standard deviation in sales. Hence, a sufficient amount of safety stock is to be maintained.
5. The owners can stay informed that festive seasons increase the revenue by 8-10%.
6. On an average, the contribution to revenue from all the weeks remained in the same for October 2022, with a festive week of Diwali included. This indicates that the

supermarket has gained a lot of regular customers and festivals influence the sales only to the slightest extent.

7. Figure 16 indicates that the supermarket does not have a specified set of dealers. They make relative decisions based on situations for placing orders. It is recommended that the supermarket increases its ties with constant dealers for timely supply.