

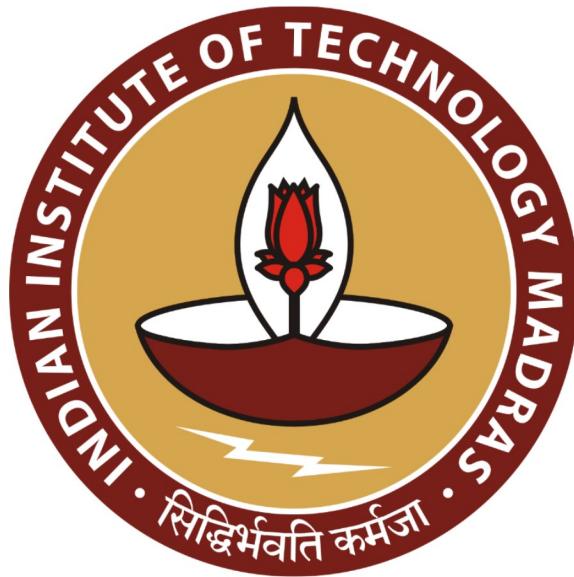
# **Enhancing Grocery Store Inventory, Customer Retention, and Profitability**

Final submission for the BDM capstone Project

Submitted by-

Name: **Yash Srivastava**

Roll number: **21F2001569**



IITM BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

## **INDEX**

Executive Summary.....	2
Detailed Explanation of Analysis Process/Method.....	2-11
Results and Findings .....	12-18
Interpretation of Results and Recommendation.....	19-20

## Executive Summary and Title

The project focuses on a grocery business “Deepak General Store” located in Vrindavan Colony Lucknow. With a focus on catering to the daily needs of nearby residents, the store operates in a business-to-consumer (B2C) model.

The store has grown significantly over time and has succeeded in building a devoted client base in the neighborhood. The store does face some obstacles, though. This project aims to analyze the available data which includes the sales data, pricing data, and inventory data in order to help the business to increase customer retention, look for inventory mismanagement if any, and ultimately enhance the organization's profitability.

Sales data will be having a major contribution to the analysis and help us judge the current situation of the business, such as the percentage of repeated vs non-repeated customers which indicates the customer loyalty of the shop. This way we can look for room for improvement present which on further analysis can help us solve the business problems mentioned above. We carefully look at the different types of factors and how they impact the business overall. Products are distributed into categories and the performance of each category separately is taken into consideration. In the end, after doing all the analysis and looking at their results, based on that the recommendations will be given to the business owner for future strategies.

## Detailed Explanation of Analysis Process/Method

The analysis begins by describing some necessary pre-processing done on the data and then we move on to the analysis.

### About the data and problem-

**For the customer-retention problem,** Customer Sales data was used for the analysis. Below are the attributes of the dataset –

Date	Customer	Item	Quantity	Category	Amount	Delivery
------	----------	------	----------	----------	--------	----------

Here, the ‘Customer’ column is the unique identifier for each customer.

This unique id is the address of the customer encoded as, building name, block number, and flat number. Example: A Customer with id N-05/F02 lives in ‘Neelgiri’ Enclave in block number 05 and Flat number F02.

Note that this was true for most customers but, some customers who were not residents of any nearby society were written their names instead. Also, there are some null values in the column for obvious reasons like the customer not willing to share his/her name, or the shop owner/worker forgetting to ask for it.

Removing such null values from the customer retention analysis is the only option we have. So, the Customer sales table is copied to another sheet where we will analyze the retention of customers. Now here we delete the rows with null values in the Customer column.

Now, as discussed earlier, there are also values in this table that correspond to rejected demand from the customer. In such a case, the ‘Amount’ value is written as -1. Since these are not actual sales done by the customer, we also remove the rows from the table with Amount with a -1 value.

**For profitability and revenue**, also we use Customer Sales data and Pricing data.

Below are the attributes of the Pricing data,

Item	Category	Selling Price	Purchase Price	Profit	Lifetime
------	----------	---------------	----------------	--------	----------

Fortunately there were no null values in this data and hardly any cleaning was required before we begin with the analysis of the profitability and revenue of the business.

In the Customer Sales data, the ‘item’ is the name of items purchased by the customer separated by commas. Similarly the ‘Quantity’ is the respective quantity of the items separated by commas. The category column is supposed to have the names of the categories in the same manner, however, there are too many null values in this column especially when the number of items is more than 1.

This will not cause any problem because the category of the item can be found using the Pricing data and hence for this reason we choose to drop this column.

**For inventory management**, apart from the ones mentioned above, we have the Inventory data which can be used for the analysis.

Below are the columns of the inventory data,

Date	Dairy and bread	Snacks	Personal Care Products	Cleaning Products	DGS namkeen	Ice-cream /beverages	Cereals	Cooking Oil	Pulses	Other food Items	Sugar	Newspaper	Water Can
------	-----------------	--------	------------------------	-------------------	-------------	----------------------	---------	-------------	--------	------------------	-------	-----------	-----------

The count left at the end of each day is recorded in this sheet and here also, fortunately, no cleaning of data is required.

## ANALYSIS

The analysis for each category of problem is written below in detail.

### **Customer Retention:**

Customer retention is the most important problem to be solved in this project. We start by looking at the Sales data we just cleaned. We have sales from May 1st, 2023 to June 30th, 2023.

There are a total of 16746 rows in this data. There are 1091 rows with ‘Amount’ equal to -1 (rejected orders), and upon removing them we get the number of sales.

Below are the basic descriptive statistical properties of this data:

- A total of 15655 sales are recorded in 61 days.
- The maximum bill amount (maximum value in the ‘Amount’ attribute) is ₹ 1550.
- The minimum bill amount (minimum value in the ‘Amount’ attribute) is ₹ 1.
- The mean bill amount (mean value in the ‘Amount’ attribute) is ₹82.

Luckily there are only a few empty rows values in the Customer attribute and after removing such rows, we are left with 14045 rows.

Now we create a pivot table of the Customer Sales data, where we keep Customer as the column and its count as unique values. We see that there are a total of 720 unique customers whose sales were recorded. Then we find minimum and maximum sales done by a customer during the period of 2 months which turns out to be 1 and 126 respectively.

To further analyze the number of orders by our customers, we create classes of the ‘Number of Orders’ of a class size of 10. Then we classify the unique customers into them based on the number of orders made by them. Example- 1-10 has ‘Count of Customers’ equal to the number of customers who made only 1 to 10 orders in the 2 months period.

Number of orders	Count of Customers
1-10	273
11-20	165
21-30	140
31-40	61
41-50	20
51-60	12
61-70	17
71-80	15
81-90	9
91-100	5
101-110	0
111-120	0
121-126	1
Total	720

From the above table, we see that the maximum number of orders of 126 by a customer is an outlier.

**This ‘Count of Customers’ based on the ‘Number of orders’ is how we wish to describe customer retention.** This makes sense because clearly, the customers who visited the shop just 1-10 times are not frequent customers of the shop whereas as we go down the table, we see the count of more loyal customers of DGS.

Now we also should consider the ‘delivery’ column in our analysis. We want to check how the delivery can impact customer loyalty and which class of customers are the ones who get their items delivered the most instead of visiting the shop.

First, make a pivot table of the Customers Sales data to figure out the count of YES vs NO in the delivery attribute. We see that 5948 orders out of 15655 are home delivered and the rest 9707 are orders done on the shop, giving us a close percentage of **38% home delivered and the rest 62% orders from the shop.**

However when the null values of ‘Customer’ are removed, then the home-delivered orders are increased to 46.6% and the rest are 53.4%

Further we try to analyze the ‘delivery’ attribute based on the number of sales done by the customers. This helps us to identify the relationship between the 2 important attributes and how delivery might be impacting customer retention.

First I filter out the rows in Customer Sales data that are home-delivered. Now I find the same simple descriptive statistical properties as mentioned above in this data.

Minimum Amount- ₹31

Maximum Amount- ₹1550

Mean Amount- ₹125

Just by looking at this we can start getting the idea that delivery has a positive impact on the amount and the sales that were made via delivery must be yielding higher revenue.

For each of the ‘Number of orders’ classes that were created in the previous table, I check the percentage of delivered orders for each customer falling in a class and take the average of those percentages of delivered orders. For example, for the range 1-10, I find all customers in this range and then calculate the percentage of delivered orders for each one of them and take their mean.

This way I can see the percentage of delivered orders for each class of ‘number of orders’ and a very clear relationship between Customer retention and delivery can be observed.

We visualize this information in the form of a graph like the previous one (shown in the Results Section) and get a clear picture of how closely Customer Retention and delivery are related.

Next, we examine the ‘Customer’ attribute of the data and try to pull out some information from it. As mentioned before, the Customer ID’s first letter refers to the housing society’s name. We can use this to further analyze customer retention of each of them. However, remember that not all the customer IDs are written this way. There are actual names of Customers as well in it. To analyze the data for each housing society, it is important to drop such rows. To check whether the value is a name or the encoded address can be tedious. For this, I checked whether the value contains special characters “/” or “-” based on which I filtered, and further checked its correctness manually by looking.

By removing the non-housing societies customers, I am left with 11678 rows which will be analyzed now.

From this, I can examine the first letter and get the society’s name. There are 5 of them from which the customers are there and from discussion meetings with the business owner, I can identify the names of the housing societies.

- N- Neelgiri
- C- Casa Greens
- H- Himalayas
- G- Govardhan
- K- Kailash

I have added now another attribute to this Customer Sales data as “Society” referring to the same. Now by creating a pivot table, let’s see how the count of customers is distributed among each of them.

Neelgiri	3970
Casa Greens	2592
Himalayas	1322
Govardhan	3094
Kailash	700
Total	11678

From the table we can see the count of customers for different housing societies. This can give us a decent understanding of how the customers of DGS are distributed in the nearby area.

### **Profitability and Revenue:**

As told earlier, for analyzing the profitability and revenue, we will be using the Customer Sales data along with the Pricing data.

To get an overview of the financial condition of the business, first, we will start by calculating the total revenue and the total profit earned by the business in the period of 2 months. This can be done using the Customer Sales data and the pricing data. Note that we take the original Customer sales data and not the one that was transformed in the previous steps for the analysis of Customer retention. The data can have Null values in the ‘Customer’ attribute for now and it will not affect the calculation of revenue or profitability.

Let's start by looking at the Customer Sales data now. As discussed earlier, the ‘Amount’ attribute in this data has some values as -1 which denotes that the customer demanded that product and it was unavailable at the shop. So it is clear that such rows do not contribute to making revenue or profit for the shop and hence can be ignored.

Hence I remove the rows with ‘Amount’ valued as -1. To calculate the total revenue earned in 61 days, I just have to sum up the ‘Amount’ values which turn out to be ₹1,283,989.

Next is to calculate the total profit. Here for every row, I must extract the individual name of items and their corresponding quantity. This can be done with Excel, however, it may require splitting values into multiple different columns, and given the size of the data, it can be very tedious. So Python is also a tool we are going to use in the analysis of Profitability and Revenue.

Our 2 data sheets will be accessed by the Python code as CSV files.

In the code, we read each line of the pricing data and store the profit margin of items in the form of a dictionary which will be used later. Then we loop over the sales data line by line and calculate the profit earned by each product by multiplying the quantity with the profit on each item and summing them all to get the total profit.

```
import csv
file= open("sales.csv")
pricesFile= csv.reader(open("pricing.csv"))
prices= dict()
for line in pricesFile:
    prices[line[0]]= int(line[4])
lines= csv.reader(file)
total=0
for line in lines:
    items= line[2].split(",")
    quantities= line[3].split(",")
    for i in range(len(items)):
        total += prices[items[i]]* int(quantities[i])
```

```
print(total)
```

Above is the code to calculate the total profit of the business in 61 days. We get the output of the code which is the total profit to be ₹251,119.

Note that this is not the net profit earned by the business owner because we are yet to consider the fixed costs such as the Electricity bill and salary of workers.

Below are the fixed costs of the business that includes Rent, electricity, Salary paid to workers, and petrol as told by the business owner. This totals up to ₹47,000.

Rent and electricity	₹8,000.00
Salary per worker	₹6,000.00
Number of workers	6
Petrol	₹3,000.00

After subtracting this cost, the net profit comes out to be:

$$\text{Net profit} = 251119 - 47000 * 2 = 157119$$

**Hence the net profit earned by the shop in 2 months is ₹1,57,119.**

Below is the list of names of the categories into which all the products present at the shop are classified.

Dairy and bread
Snacks
Personal Care Products
Cleaning Products
DGS namkeen
Ice-cream /beverages
Cereals
Cooking Oil
Pulses
Other food Items
Sugar
Newspaper
Water Can

Now we want to look at the revenue as well as profit made by each of these categories in order to judge their individual contribution to the shop's profit.

For a similar reason, we will be using Python for finding this.

```
file= open("data.csv")
pricesFile= csv.reader(open("pricing.csv"))
profit= dict()
category= dict()
sellingPrice= dict()
cats= set()
for line in pricesFile:
    profit[line[0]]= int(line[4])
    category[line[0]]= line[1]
    sellingPrice[line[0]]= line[2]
    cats.add(line[1])
lines= csv.reader(file)

catRevenue, catProfit= {}, {}
for i in cats:
    catRevenue[i]= 0
    catProfit[i]= 0

for line in lines:
    items= line[0].split(",")
    quantities= line[1].split(",")
    for i in range(len(items)):
        catRevenue[category[items[i]]] += sellingPrice[items[i]] *
int(quantities[i])
        catProfit[category[items[i]]] += profit[items[i]] *
int(quantities[i])
print(catRevenue)
print(catProfit)
```

The above code prints out the revenue and profit earned by each category in form of a dictionary which is manually copied into the Excel sheet.

These values are now used in plotting two different graphs, Category Vs Revenue and Category Vs Profit (shown in the Results section).

The reason for doing this is that when we take a closer look at these graphs, we gain a clear understanding of how each category plays a crucial role in impacting the financial health of the

business. This insight allows us to determine how we can utilize this information to make changes that will help enhance the overall profitability of the business.

### **Inventory Management:**

Another aim of the project was to look for mismanagement in inventory and give insights on its improvement if any. Below is data found using a pivot table on pricing data to get the number of types of different SKUs in each category.

Category	Types of different SKUs
Dairy and bread	44
Snacks	109
Personal Care Products	72
Cleaning Products	35
DGS namkeen	22
Ice-cream /beverages	65
Cereals	12
Cooking Oil	8
Pulses	6
Other food Items	32
Sugar	1
Newspaper	3
Water Can	1

There are a total of 13 categories and 410 different products. Out of these most products fall under Snacks. On the other hand, ‘Sugar’ and ‘Water Can’ are single products.

As discussed before, in the sales data some of the rows have their Amount valued as -1 which signifies that the item is unavailable at the shop and cannot be provided to the customer. Note that these items demanded by customers are relevant to the domain of the shop and are the ones that should have been available at the shop in an ideal case.

So first we start by filtering out the rows from the Customer Sales data that have the attribute ‘Amount’ as -1. This is now the data to be analyzed to look at the rejection/unavailability of items in the shop. For most of these rows, the ‘Quantity’ attribute is missing which is fairly justifiable as the customer did not tell us the quantity they need if the item is already out of stock. We can simply assume the unit quantity for each of such rows.

Let's start with calculating the apparent revenue and profit of these rejected orders.

We use the VLOOKUP function in Excel to look at the Selling Price of the item and then we sum up all the prices giving us the revenue.

Similarly, we find the profit by again using the VLOOKUP function by getting the profit of each of the items and then all of them summed up.

**The values for revenue and profit are calculated to be ₹1,83,989 and the profit is ₹23,455.**

These are the loss in revenue/profit, due to rejected orders in 2 months.

After this, we must visualize the number of orders rejected from each category of products.

### **Is the data accurate?**

When we talk about the Customer Sales data, the data is been collected at both the shops by the business owner and the workers. As much as the owner has tried to keep the data up to date, there are chances that there can be some orders that can be missed to be noted down, especially during rush hours. The owner conveyed the same message to me. This can cause a little inaccuracy in the analysis, however, the analysis should still be really insightful. This is because we still get more than 15 Thousand rows in our sales data which is sufficient to see the trends and patterns.

Also, as discussed before, the 'Customer' attribute which is the ID for each customer has a few missing values which could be because either the customer refused to share their name or the Owner/worker could not note it down due to hurry.

On the other hand, the pricing data is a one-time noted data and there should not be any inaccuracy in its data. Similarly, the inventory data is noted once each day and it can be assumed that the data is fair and accurate.

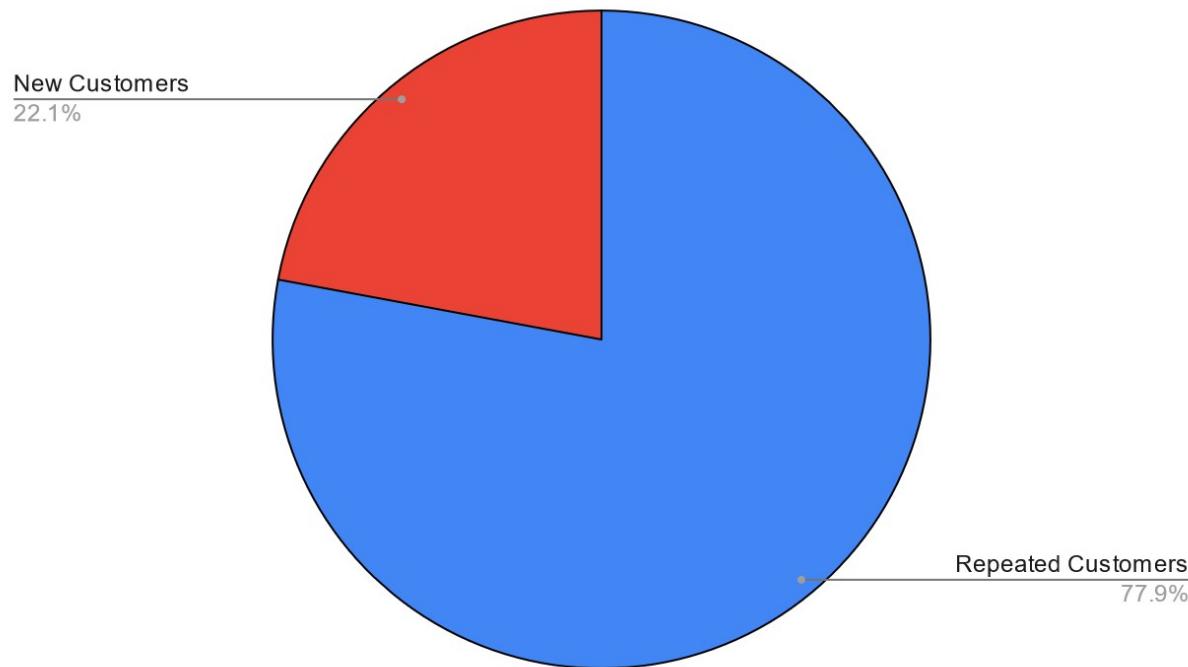
## Results and Findings

Below are the results and insights gained so far from the analysis.

Starting from the **Customer Retention** problem, below are the descriptive statistics of the Customer Sales data summarized in the form of a table.

Total no. of Sales	15655
Mean sale per day	256
Max bill amount	1550
Min bill amount	1
Mean bill amount	82
Min sales by a customer	1
Most sales by a customer	126
Mean sales by a customer	23.33

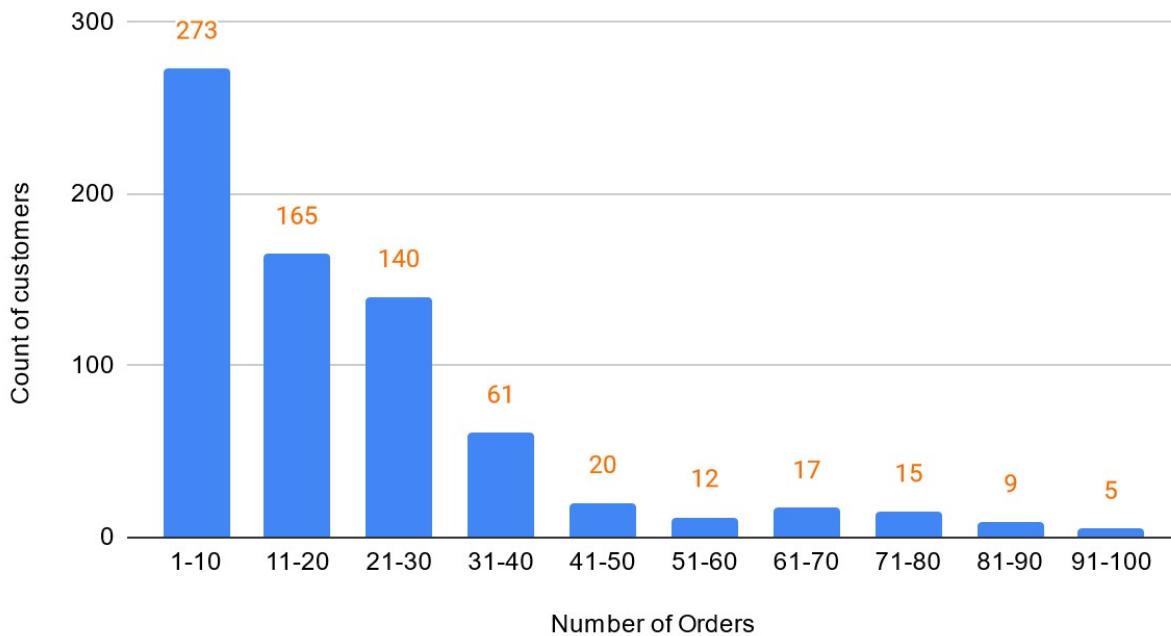
The number of sales by a single customer lies between 1 and 126 with a mean value of 23.33



Pie Chart showing Repeated customers

The Pie chart shown above determines the count of customers who just visited the shop once. There are 720 unique customers who visited the shop in a duration of 2 months out of which **561** customers visited the shop at least twice and the rest **159** visited the shop only once.

### Count of customers vs. Number of Orders

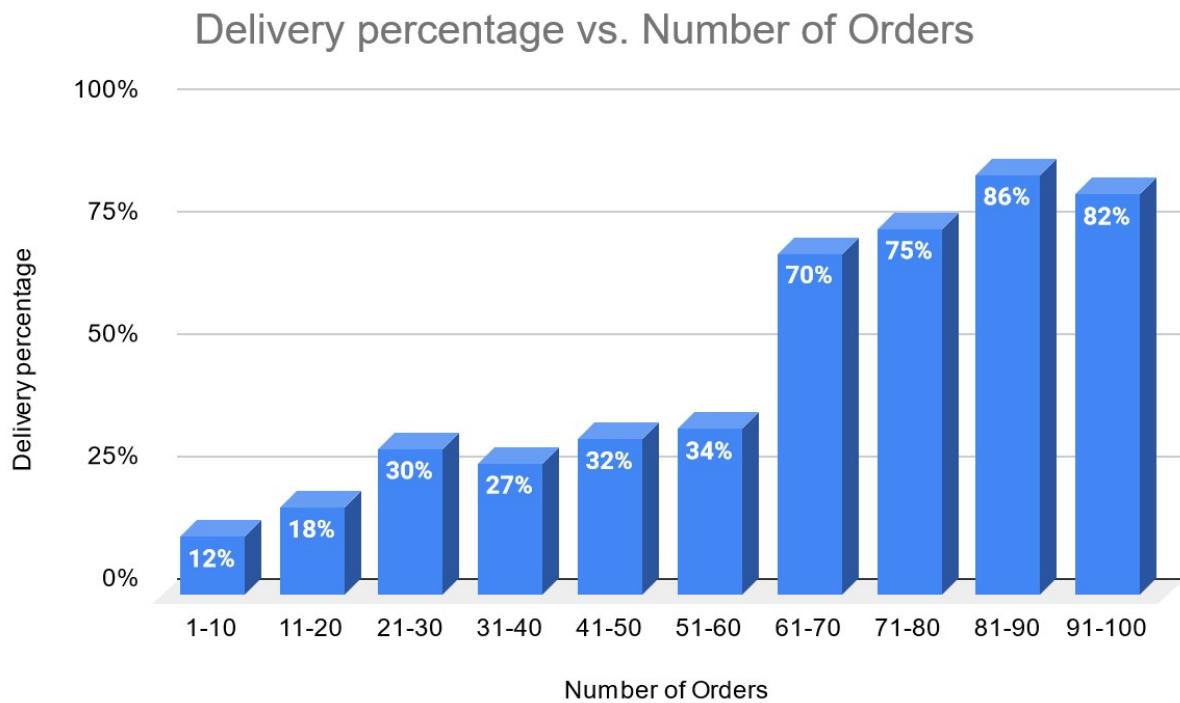


The above graph shows the count of customers based on the class of number of orders done from the shop. We observe that the maximum number of customers have ordered a mere **1-10** times from the shop. This is also because of the high number of customers who visited the shop just once during the period of 2 months.

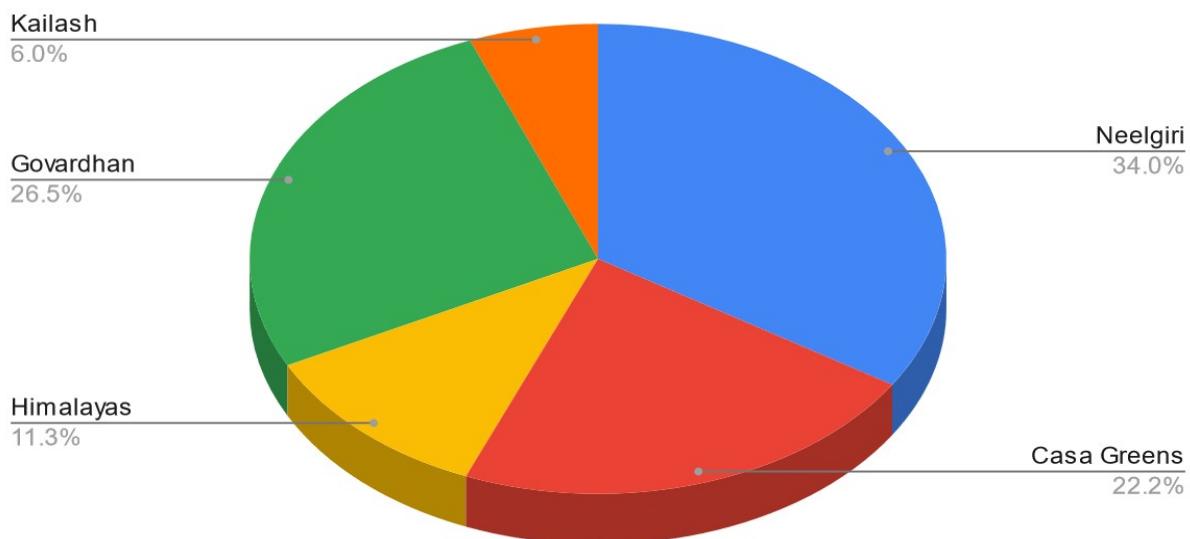
Earlier from the Pie Chart, the Customer Retention of almost 78% seemed fine, but here we get the actual pictures. Most of the customers, who visited the shop more than once are also not frequent customers of the shop.

Given the fact that the shop supplies the daily needs of the customers such as Dairy, Water Cans, and newspaper, it is expected that there is much higher customer retention than at present.

Looking at the role of delivery on Customer Retention was an important part of the analysis. Below is the graph that throws more light on it by visualizing the average percentage of delivery orders by the customers.



From the above graph, we can see the strong impact of delivery on the number of orders by a customer. We can see that the most loyal customers prefer delivery as their option. **A significant rise can be seen in the delivery percentage for the customers who ordered from the shop more than 61 times in 2 months.** This helps us conclude that customers who frequently use the option of delivery are more loyal customers to the shop and hence the **free delivery is a great way to promote customer retention.**

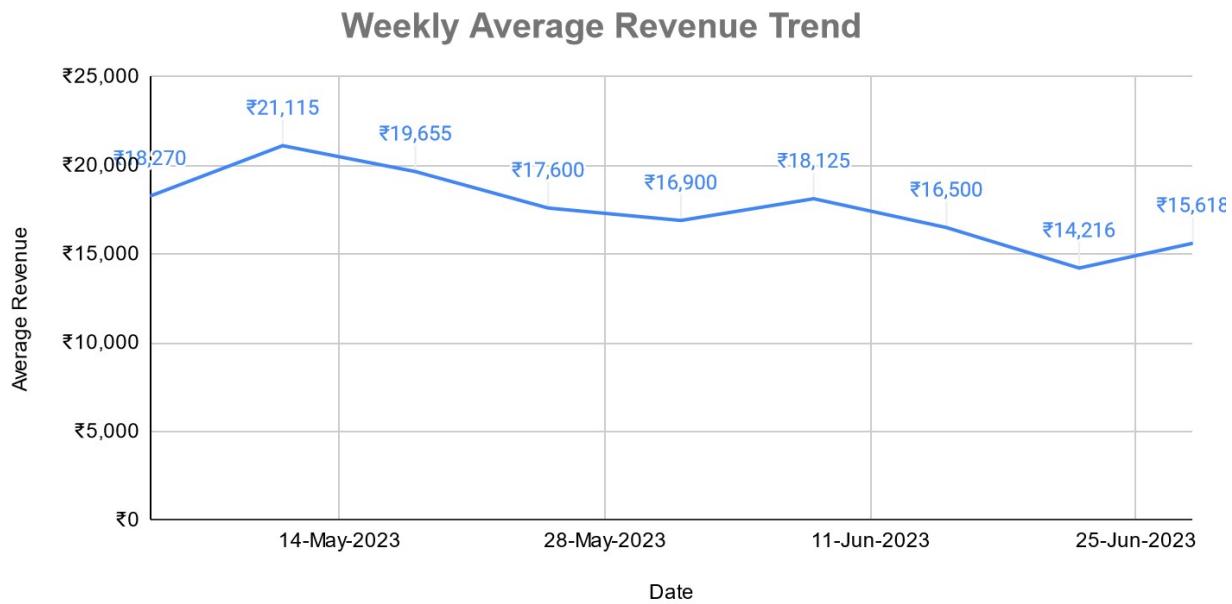


Pie Chart shows the contribution to Sales into different societies

Now, in the above Pie Chart we are looking at the contribution to sales for different housing societies around DGS.

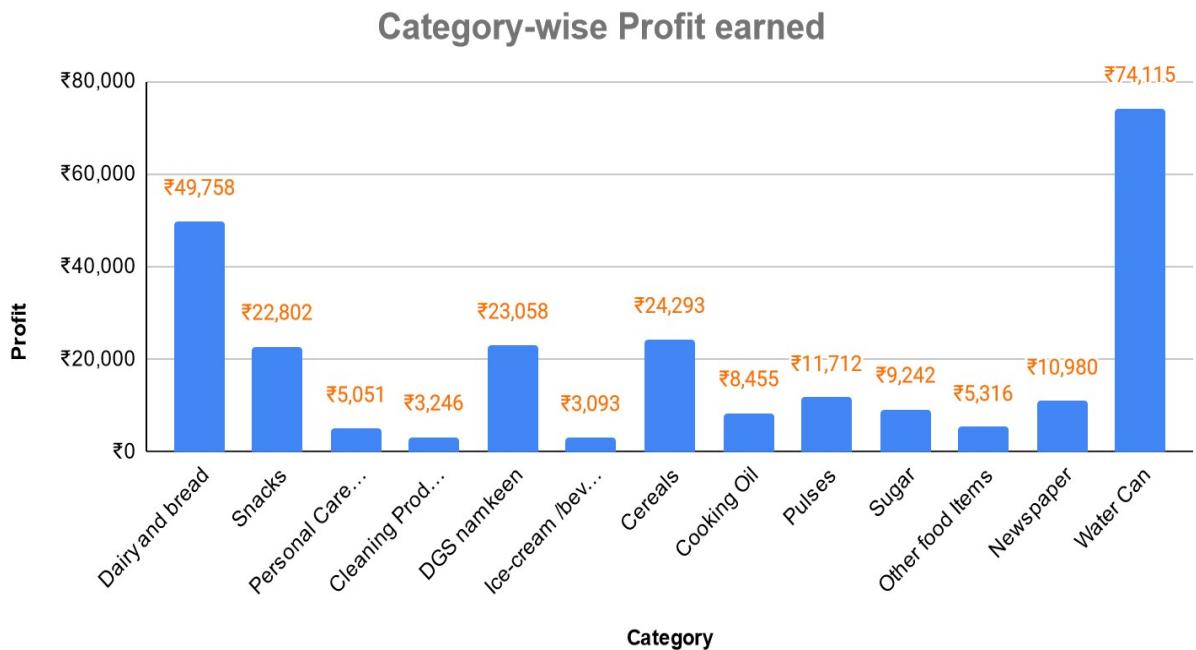
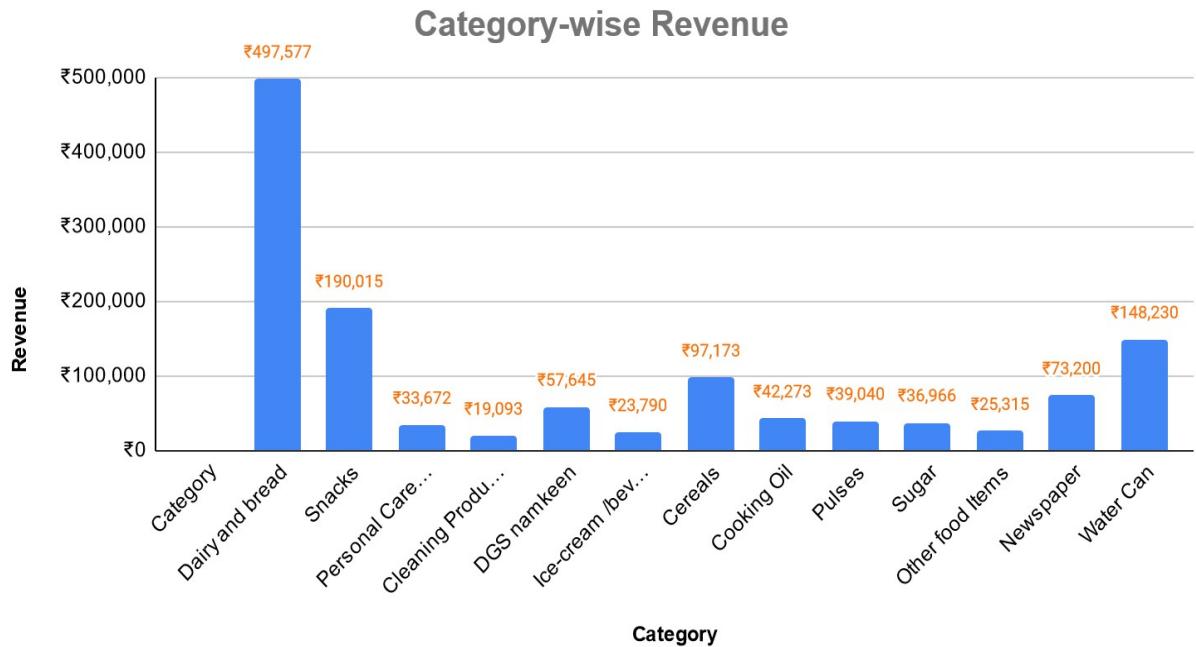
All the 5 housing societies mentioned above are within a range of 1.5 kilometers from either of the shops. ‘Kailash’ and ‘Himalayas’ are contributing less to the sales of DGS and some actions must be taken to increase sales from these societies. **Also to note that the Himalayas Enclave society is the biggest and most populated of all (roughly populated equal to the sum of the rest of all) yet has a mere 11.3% contribution to the sales in DGS.**

Next is to look at the trend of the daily revenue, below is a graph where the average weekly revenue for the period of 2 months (simply made by making a pivot table of sales data),



From the graph, the relative decrease in revenue is seen from the start of June month. However, this is not a reason to be worried since this was a period of heavy rainfall. For some days, it was raining all day and fewer customers visited the shop during such time.

Hence it can conclude that the daily revenue of the business is not volatile.

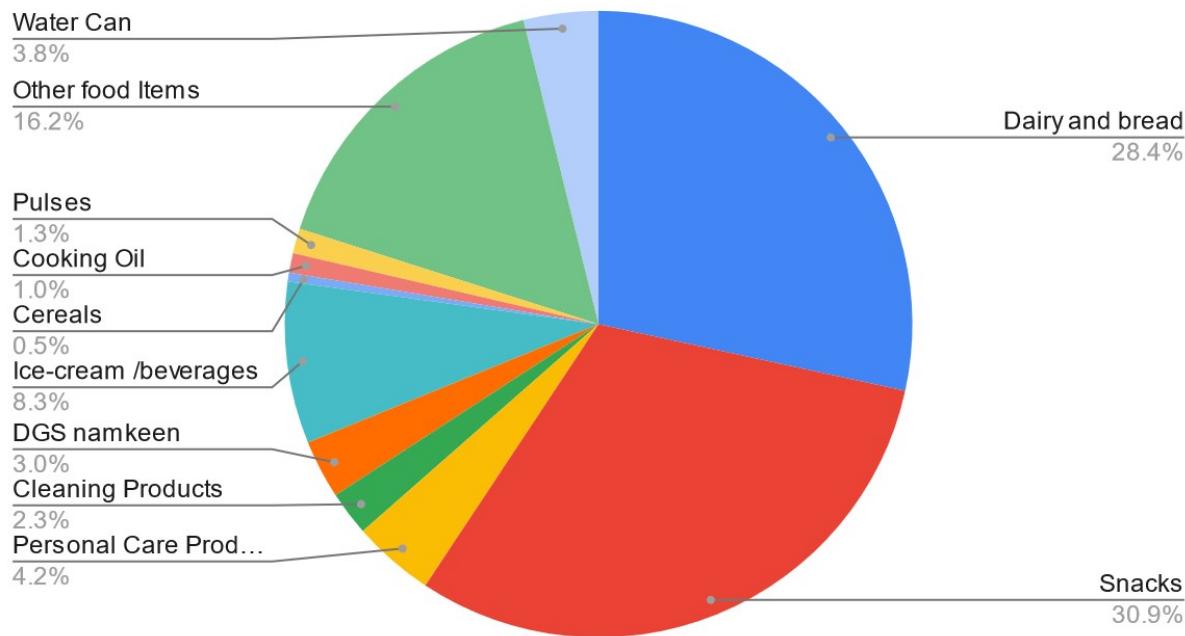


Above are the 2 graphs that display the Total Revenue and Total Profit earned by each category respectively. We can see **Dairy and Bread category has earned the highest revenue (₹497,577.00)** for the business whereas **Cleaning Products has earned the lowest revenue (₹19,093.00)**

The 2nd and 3rd revenue is made by Snacks and Water Can. This denotes that the customers look at the shop as a medium to fulfill their daily needs for which they do not wish to go far away. For categories such as Personal Care Products, and Cleaning Products, customers' first choice is not DGS, they could be ordering online or buying stuff from some main market.

From the next graph for Total Profit, we see that though Dairy and Bread has the highest revenue, due to a relatively lower profit margin it is not the highest profit earner for the shop. **Water Can** has generated the most profit in 2 months due to its high-profit margin of nearly 50%. **Out of the 13 categories, Water Can alone generated more than one-third of the total profit for the business.** Also to mention that Water Can category is just one single product, thus promoting Water Can will be highly beneficial for the shop.

**Number of rejected orders based on Category**



**Pie Chart of distribution of rejected orders based on categories**

The above Pie chart compares and shows that close to 60% of rejected orders belong to the Category of 'Dairy and Bread' and 'Snacks'. During multiple discussion meetings with the business owner, it was revealed that he deliberately tries to keep the number of Milk packets for everyday sales a little lower than expected. This strategy is maintained because milk items have a lifetime of only 1 or 2 days, and avoiding wastage is prioritized over avoiding out-of-stock before time. This is justified by the low-profit margin in Milk products and his estimate from

experience says that the loss from 1 unsold SKU passing its expiry date is equivalent to the profit earned from 10 such SKUs being sold.

On the other hand, **'Snacks' and 'Other Food items' have a higher lifetime and their SKUs can be avoided going out of stock.**

Category	Net Profit Margin (in percent)
Dairy and bread	10
Snacks	12
Personal Care Products	15
Cleaning Products	17
DGS namkeen	40
Ice-cream /beverages	13
Cereals	25
Cooking Oil	20
Pulses	30
Other food Items	21
Sugar	25
Newspaper	15
Water Can	50

Above is a table of the approximate net profit margin for each category of product, which was simply calculated using the using the pricing data. We can verify that the profit margin is maximum for Water Can at 50% and is followed by DGS Namkeen with 40% which is a personal brand for the shop. If sales of such items are boosted can earn a lot of profit for the business.

## **Interpretation of Results and Recommendation**

Now that the analysis and the findings have been looked at, the report can be concluded with the interpretation of results and recommendations that would help the business to move forward for its benefit.

The following are the interpretations along with recommendations

### **1. Avoiding stock-outs of products-**

From the Pie chart “Number of rejected orders based on Category” showing the percentage of rejected orders we saw that categories like ‘Snacks’ are also facing out-of-stock outs despite being a product category that has a relatively higher lifetime and is also a category that is forever in demand. Stock-out is something that not only decreases the revenue of the business but more importantly, it affects the long-term Customer retention of the shop, hence more stocks of Snacks must be gathered in the inventory.

### **2. High emphasis on free delivery-**

Another important result seen from the “Delivery Percentage Vs Number of orders” graphs was about Customer Retention and how delivery impacts it positively. The shop must promote the free delivery service and make people aware that it provides. It is quite natural that once the option of free delivery is shown to people, they will tend to make DGS as their first choice for ordering any shop/service. This will also encourage people to buy other products from DGS and eventually, the sales and revenue of other products such as Personal Care Products and Cleaning Products will increase. People nowadays are active users of other grocery delivery services such as Blinkit, Big-Basket, and Amazon where they have to pay between ₹10 to ₹50 as the delivery charge for their order. If such a service with no delivery cost is provided by DGS along with the liberty to pay on delivery to the various resident societies then it is going to boost Customer Retention, revenue, and profitability.

### **3. Target Promotion on Housing Societies -**

As inferred from the Pie chart for distribution of sales across different societies, we saw that despite the huge size and Population, the residents of ‘Himalayas’ society have very few orders from DGS. Pamphlets should be distributed outside these societies and make people more aware of DGS and its free delivery service. Given the size of this society, it would make a huge impact on the revenue and profitability of the business. Also, they should present themselves more as a home-service provider than a shop making them peoples’ first choice for ordering items online.

**4. Promoting high-profit margin items-**

We saw from the “Category wise profit earned” graph, that how Water Can and DGS Namkeen has relatively higher profit margin than others. Some offers should be given to such products in order to attract more sales for these products. These categories can contribute a lot to the profit of the shop. For DGS Namkeen which is a personal brand, a better packaging and quality check should be done and the shop owner should recommend consuming this to customers over other Snacks. Other marketing strategies such as placing them in the front row of the shop where the customers can notice it quickly must be followed.

5. The shop stores data in notebooks and is not maintained properly. The shop owner owns a computer which is kept in the new shop. However, the computer is never used for any purpose. They should try to start storing the data into Excel sheets over notebooks for various reasons for which I am ready to help them on a personal level. This will help in keeping the data more accurate and easy to analyze. Further with time they can also purchase a code scanner machine for products (which on a quick search can be found not costing around ₹2000 to ₹3000) and adapting a better technology culture.