IIT Madras

B.S. (Programming and Data Science)

BDM capstone project mid-term report

on topic:

# DRIVING SALES GROWTH AND MAXIMIZING PROFITS FOR


Submitted by:

**Mohak Khatri**

**21f1000865**

# INDEX

# EXECUTIVE SUMMARY

In this mid-term report, I've provided an overview of my business data management project conducted to maximize profits and drive sales growth for a small achar-murabba storefront. In this project, I have focused on analyzing data from various sources, including handwritten purchase records, scanned printed bills, and a customer preference survey.

The initial analysis of the customer preference survey revealed that the most popular product categories are "Roasted/Baked packed snacks" and "Spices," while "Murabbas" are in a lower demand. Quality of the product is the most important factor for the customers. Products with a high quality would attract almost all customers. For edible products, quality and flavor act as the most desirous features of the product.

I will determine the quantity of the products (of a specific category) sold in a particular month by combining the data of the "Purchase_data" and "Company_category_mapping" datasets. This would help me in understanding the shortcomings of the inventory of the shop as well as help me to find an optimal product range to be selected for sale. I'm planning to do this detailed analysis in the end term report.

This mid-term report highlights the progress made in data collection, analysis, and preliminary findings. The final report will focus on detailed analysis and possible recommendations would be given to the business to drive their sales growth and maximize their profit.

# PROOF OF ORIGINALITY OF DATA

I am sharing the google drive link which contains the owner interaction video, the authorization letter handover images and the image of the data-authenticity letter (stamped and signed by the owner).

## PRIMARY DATA

Although I couldn't get the sales data due to no digital records, I collected the purchase data from the shop (as advised by the instructors in a live session). The link to the written records pdf folder is:

[redacted]

I have also conducted a customer preference survey through google forms. The link for the same is mentioned below:

https://docs.google.com/forms/d/e/1FAIpQLSdMaZEssu6nJ1kWmpEPjvpbMRsbUH_gjbkCkm1wJ-yEKQ_5xQ/viewform?usp=sf_link

The responses of the survey are collected in the following spreadsheet:

https://docs.google.com/spreadsheets/d/1pENUJ9jw9hvGQvCHSjiebXHNdjN85bAFE2o4SKtaT2g/edit?usp=sharing

I had also collected and scanned whatever printed bills were available with the owner (only for the months of April and May). I cannot consider the bills for the month of April because some of the bills were missing and hence would lead to erroneous data. The link of the pdf of available bills is mentioned below:

[redacted]

# DATA COLLECTION PROCESS

I had a meeting with the owner regarding the intended data collection. I told him that I would require sales data and purchase data from him. He told me that he won't be able to give the sales data as they do not produce printed bills for customers. Then, as per the advice of the instructor in the live session, I asked the owner for just sales data. The owner told me that printed bills were only available for the month of April 2023 and May 2023, they discard the rest of the bills. He told me that he creates written records for the total amount paid to a company month-wise. So, he provided me with the month-wise payment data i.e., the total amount of money paid, for past 6 months. Almost all of these companies have a single category of product and I will be able to analyse the popularity/demand of certain products.

Hence, from the shop, I have the month-wise total billing and some of the printed bills for the months of April 2023 and May 2023.

I had also intended to collect customer preference data and I have done that through a customer survey using google forms. I have had the form circulated within my city to people of various age groups. This will help me better understand the customer preferences. Both of these datasets are being used to perform inventory and customer preference analysis.

## DATA DIGITIZATION

The main purchase data was available only in handwritten pages. I have converted each month's data into excel spreadsheets. The link to that spreadsheet folder is mentioned below:

[redacted]

I had scanned and recorded a pdf of all of whatever printed bills were available and I have manually extracted the company-category-price mapping and stored it in excel. I wasn't able to find the categories of some companies, these companies' cells are formatted with red colour in the sheet. I also tried to confirm the left-out companies with the owner and he told me whatever he could gather but still some companies are left which I have left for the scope of end term report. The link to the mapping sheet is mentioned below:

[redacted]

# METADATA AND DESCRIPTIVE STATISTICS

## METADATA

Groups of data/terminology used:

1. Non_Digitized_Data: Purchase_data, Printed_bills
2. Digitized_Data: Purchase_data, Company_category_mapping
3. Survey_Data: Customer_survey_responses

**Non_Digitized_Data metadata:**

1. Purchase_data metadata:

   There are total 6 files, each file represents the purchase data of that particular month. Example: "oct22.pdf " has the scanned image of the handwritten page of the data for the month of October, 2022. Each handwritten Page has 4 columns.

   The first column stores the name of the company along with its GST number. The first column itself stores the dates on which any payment has been done to the company. Rest of the 3 columns, labelled as "5%", "12%" and "18%" are used to store the total amount (in INR) paid to the company at the respective date from column 1 according to the respective GST ranges. E.g.: if column 1 stores "abc company", "12oct22" and this row has an entry in the "5%" header column as "3344" this means that "[redacted] has paid INR 3344 (+5% GST) to "abc company" on 12th of October 2022.

2. Printed_bills:

   The format of every company's printed bill is different however there are some common features that every bill has which are mentioned as follows:

   2.1. Company name (biller company)

   2.2. Client name (billed-to company, "Jain Pickle Wala")

   2.3. Name of the item

   2.4. Date of invoice

   2.5. GST slab

   2.6. Quantity purchased

   2.7. Price per item

   2.8. MRP of the product

   2.9. Total price of the item

   2.10. Grand total of the amount to be paid

**Digitized_Data metadata:**

1. Purchase_Data:

   Contains 6 spreadsheets, each of which has almost the same structure as its handwritten counterpart, the only difference being the splitting of "name" of company under the column header "company" and the "date" of purchase under the header "date".

The purchase_data folder also contains company-category mapping spreadsheet described separately below.

2. Company_category_mapping:
   Spreadsheet named "company_cat_mapping" contains 4 main columns and 1 auxiliary column. The main columns are:
   2.1. "Company": Stores each company's name from the purchase data
   2.2. "category": Stores the category of products purchased by the shop (from the respective company)
   2.3. "gst rate": Stores what slab of GST the specified category products from the company fall under
   2.4. "price_range": Stores the effective price that the shop has to pay to acquire the one product from that particular category of that particular company

   Auxiliary column: stores the names of those companies whose category mapping is yet to be done.

**Survey_Data Metadata:**

Customer_survey_responses:
This spreadsheet contains the questions as columns. The timestamp at which a particular response has been submitted is stored under the header "Timestamp". The spreadsheet contains 98 responses at this moment. Any future responses would be appended in the same spreadsheet as further rows.

# DESCRIPTIVE STATISTICS

**Company_category_mapping:**
Maximum, minimum value of price range: INR 500, INR 3

**Purchase_data:**

Due to the space constraints, the descriptive statistics (Mean, Median, Min, Max, Variance and Standard Deviation) have been recorded in the spreadsheets itself.

# ANALYSIS PROCESS/METHOD

Since I have digitised the purchase data into Google Sheets, I will use basic spreadsheet features and tools to perform the data analysis.

For now, I have used the analysis off the customer preference survey data wherein the analysis was provided by Google forms. Basic bar graphs and pie charts were produced for each question which has helped me in understanding the trend of the customer preferences.

I've also planned to use a standard library for handling excel files and spreadsheets into Python, which would help me in analysing the data faster and in a more efficient way. Since I have to club data from several sheets and then perform the required analysis doing it with the help of Python would give me much more flexibility to use custom code according to the complex requirements.

I am also planning to use the Pandas library to handle the data and produce it into data frames which is yet another industry standard.

Currently I have performed the analysis of the "Customer_survey_responses" data and I have discussed about the customer preferences with the owner.

I have reserved the detailed analysis of the purchase data for the scope of final report submission of this project.

Detailed method of analysis that I will use is described as follows, I will determine the quantity of the products (of a specific category) sold in a particular month by combining the data of the "Purchase_data" and "Company_category_mapping" datasets. This would help me in understanding the shortcomings of the inventory of the shop as well as help me to find an optimal product range to be selected for sale.

Since I already have the customer preference data from the survey and the respective analysis for the questions, I've tried to find the most purchased categories by the customers. Also, I am trying to find what all factors lead to a product being more popular than other products of the same category. I am also trying to find the factors responsible for making a customer choose a particular shop over another.

# RESULTS AND FINDINGS

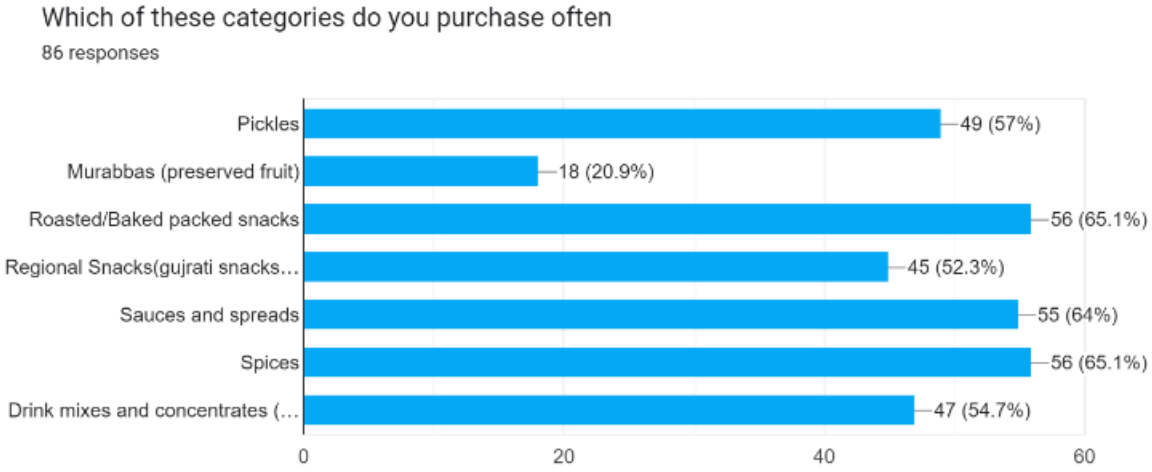


*Fig. 1 "category_popularity"*

There are 2 most purchased categories: "Roasted/Baked packed snacks" and "Spices".

"Murabbas" (preserved fruit) which is a very specific product of the business, is the least purchased product, i.e., only very specific customers are likely to buy it and it should account for least possible holding of cashflow as well as inventory.

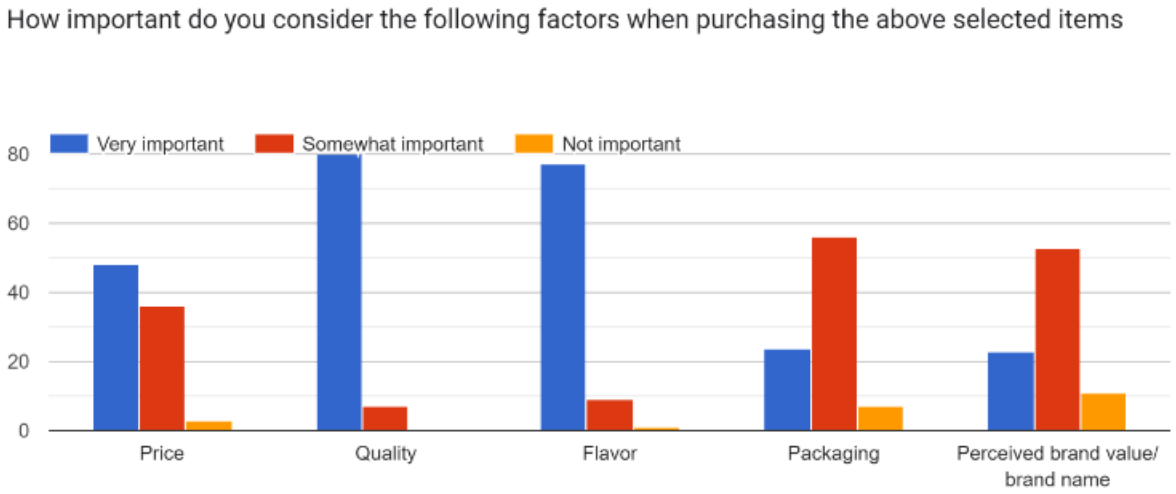


*Fig. 2 "Factors_deciding_product_popularity"*

The above graph clearly shows that quality of the product is the most important factor for the customers. Products with a high quality would attract almost all customers. For edible products, quality and flavor act as the most desirous features of the product.

The following graph shows that the majority of the population is willing to pay more for organic products.

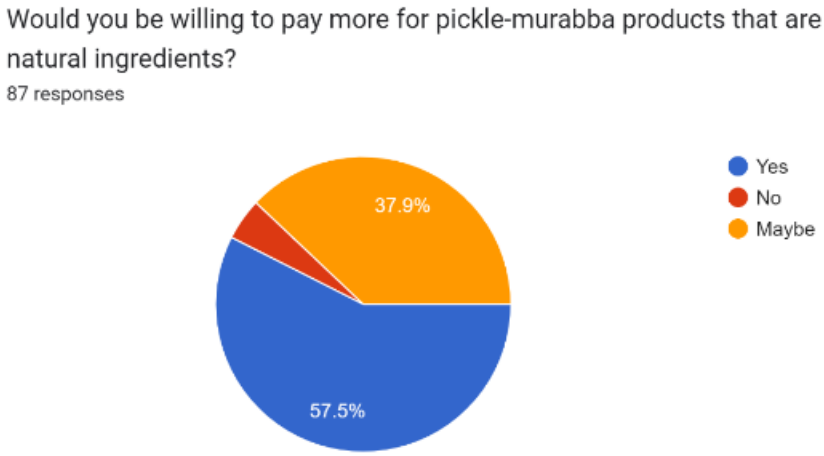


*Fig. 3 "Organic_products_demand"*

## THANK YOU