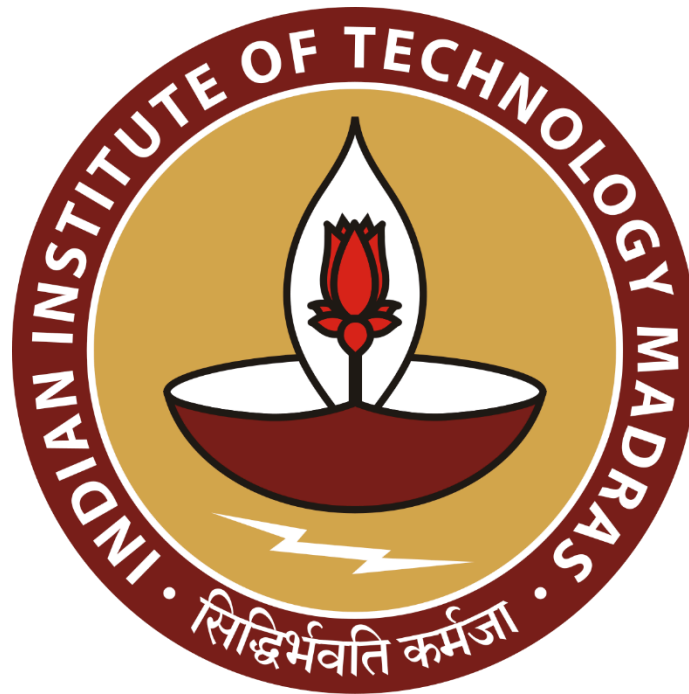# Sales Trends and Customer Behavior in a Medical Shop: An Analytical Study

Final Submission for the BDM capstone Project

Submitted by

Name: SOUMYADEEP DHALI

Roll number: 22f1001762

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

1

Contents:

1. Executive Summary

The data-driven analysis for DEY MEDICO provided actionable insights to boost sales, improve customer loyalty, and optimize inventory management. Through RFM and Customer Lifetime Value (CLV) analyses, the shop's most valuable customers were identified, allowing for targeted retention efforts such as personalized discounts, loyalty rewards, and early access to promotions for top spenders like Rupam Dasgupta and Sujata Bose. The findings also highlighted a need to nurture potential loyalists and re-engage at-risk customers with tailored marketing strategies.

Despite a relatively low churn rate, there is potential to further strengthen customer retention by introducing loyalty programs, enhancing customer service, and implementing competitive pricing strategies, especially in response to neighboring competitors like Apollo Pharmacy. Additionally, the correlation analysis between RFM dimensions suggested that increasing purchase frequency could lead to higher overall spending, emphasizing the importance of encouraging repeat purchases.

Inventory analysis pinpointed peak sales months, such as June and July, suggesting a strategic approach to stocking inventory during high-demand periods to maximize sales and minimize the risk of stockouts. The data also revealed opportunities to optimize inventory levels by analyzing historical sales trends and adjusting stock accordingly to prevent overstocking or stock outs during both peak and off-peak seasons.

By leveraging these insights, DEY MEDICO can enhance customer satisfaction, streamline operations, and ultimately drive higher profitability. This comprehensive approach to data-driven decision-making will not only fortify DEY MEDICO's market position but also ensure sustained growth in an increasingly competitive retail environment.

Tools Used:

Pandas: For data manipulation and analysis with DataFrames.
Excel: For organizing, analyzing, and visualizing data in spreadsheets.
Matplotlib: For creating static charts and plots.
Seaborn: For making attractive statistical graphics easily.

2. Detailed Explanation of Analysis Process/Method

Process of Data Analysis in Sales Optimization:

The journey of data analysis comprises several crucial steps, each contributing to the overall understanding of the data. This encompasses defining the problem, collecting and organizing the data, data cleaning, transforming it, doing analysis, and ultimately drawing meaningful conclusions.

Defining the Problem

Objective:

The primary objective is to improve the sales performance of DEY MEDICO, a medical shop. This involves identifying opportunities to enhance revenue, increase customer loyalty, and optimize operational efficiency through data-driven strategies.

Tasks:

Analyze Current Sales Performance: Review the current sales data to identify baseline performance metrics. Understand the product mix, customer demographics, peak sales periods, and any existing promotional strategies.

Identify Pain Points: Uncover specific challenges DEY MEDICO is facing, such as declining sales, customer attrition, or competition. Determine if there are particular product categories or customer segments that need attention.

Develop Hypotheses: Based on the initial understanding, formulate hypotheses that will guide the analysis. For instance, "Introducing targeted promotions for high-margin products will increase overall profitability" or "A rewards program will improve customer retention.

Collecting and Organizing the data

In order to gather the sales data for this analysis, significant effort was required. After being declined by seven other medical shops due to their strict confidentiality policies, I successfully established a trusting relationship with DEY MEDICO, which agreed to share their one-year sales data with me. Building this trust was essential to secure the data collection, as confidentiality was a key concern for the shop.

Given the uniqueness of the situation, I encountered challenges in clearly articulating how I could support the business. This uncertainty also impacted my ability to define specific data requirements that could generate actionable insights. The ambiguity limited my confidence in explaining how my analysis might contribute to improving their operations. Many stakeholders felt they understood their needs but faced complexities in practice that data analysis alone couldn't easily resolve. Skepticism arose, with many doubting that data analysis would reveal anything new or valuable.

In light of these obstacles, I am dedicated to refining my approach. I plan to allocate more time to relationship-building and fostering mutual understanding. This includes deeply engaging with the businesses, showing genuine interest, and gradually clarifying how data insights can inform better decision-making. By maintaining regular communication and adopting an empathetic approach, I aim to build trust, which is essential for successful data acquisition and analysis.

Ultimately, I succeeded in obtaining data from a willing business owner.

Data Cleaning

DATE	BILL NO.	PARTY NAME	BILL AMT.	TAXABLE	TAX	TAX FREE	EXEMPTED	R. OFF

These were the available features in the sales data provided by the owner. It consisted of a blend of clean and mixed information which required intensive cleaning before I could get started with my analysis.

The data consisted of 1243 rows and 9 columns. Subtotals were present for each day in the data which was tampered with my analysis. So to achieve uniformity, I cleaned fields with inconsistent formatting in the data.

For instance, the 'TAX FREE' and 'EXEMPTED' columns were recorded in the dataset, but all the values were zero, rendering them redundant. As a result, I

decided to drop these columns from the dataset. Additionally, there were some rows containing NaN values that needed to be addressed.

For my Seasonal Analysis i categorized the months broadly into three categories i.e, Spring, Summer, Winter.

These categories have been elaborated upon in the mid-term report.

RFM Analysis

At first I did a RFM analysis to better understand customer behavior and segment the customer base based on their purchasing patterns. The goal is to identify key customer segments and tailor marketing, sales, and retention strategies accordingly.

RFM analysis is a marketing technique used to evaluate and segment customers based on their purchasing behavior. It focuses on three key metrics:

1. **Recency (R):** How recently a customer has made a purchase. More recent purchases suggest higher engagement.

2. **Frequency (F):** How often a customer makes a purchase. More frequent purchases indicate greater loyalty.
3. **Monetary (M):** How much money a customer spends. Higher spending reflects a more valuable customer.

By analyzing these metrics, businesses can segment their customers into categories such as "Champions," "Loyal Customers," "At Risk," etc. This segmentation helps tailor marketing strategies to target different customer groups effectively, aiming to improve retention, increase sales, and optimize marketing efforts.

Detailed Steps for my RFM analysis:

1. Converted the 'DATE' column to datetime format using the specified date format (DD-MM-YYYY). Any errors in conversion will result in NaT (Not a Time) values.

2. Removed rows where any of the specified columns ('DATE', 'PARTY NAME', 'BILL AMT.') have missing values.

3. Calculates RFM (Recency, Frequency, Monetary) metrics:
   ● Recency: The number of days since the last purchase.
   ● Frequency: The total number of purchases.
   ● Monetary: The total amount spent.
Groups by 'PARTY NAME' and aggregates the metrics

4. Renames columns in the rfm_table DataFrame for clarity.

5. Recency: Scores are assigned based on quantiles, with 1 being the most recent and 5 being the least recent.

6. Frequency and Monetary: Scores are assigned using custom bins.

7. RFM Score: The sum of Recency, Frequency, and Monetary scores.

8. Prints the first few rows of the RFM table to the console.

9. Saves the RFM table to a CSV file named 'rfm_results.csv'.

10. Defined a function to segment customers based on their RFM scores.

11. Applied the segmentation function to each row in the RFM table.

12. Printed the first few rows of the segmented RFM table to the console.

13. Saved the segmented RFM table to a CSV file named 'rfm_segmented_results.csv'.

14. Printed the bins used for scoring Frequency and Monetary metrics.

15. Count Segments: Counts the number of customers in each segment.

16. Plot Distribution: Creates a bar plot to visualize the distribution of customer segments.

The results and finding of this has been given in *'Results, Findings, and Interpretation of Results'* section

Customer Lifetime Value(CLV) analysis:

Customer Lifetime Value (CLV) measures the total revenue a customer is expected to generate during their relationship with a business. It helps businesses:

- **Prioritize high-value customers**
- **Allocate marketing budgets effectively**
- **Enhance customer retention strategies**

CLV provides insights for optimizing growth and profitability.

Detailed Steps for my Customer Lifetime Value(CLV) analysis:

```
df['CLV'] = df.groupby('PARTY NAME')['BILL AMT.'].transform('sum')
```

df.groupby('PARTY NAME')['BILL AMT.']: This groups the DataFrame df by the 'PARTY NAME' column. This means that all rows with the same 'PARTY NAME' are grouped together.

transform('sum'): For each group, it calculates the sum of 'BILL AMT.' and assigns this sum to every row in the group. The result is a new column 'CLV' where each row contains the total billing amount for the corresponding customer.

```
clv_summary = df[['PARTY NAME',
'CLV']].drop_duplicates().sort_values(by='CLV', ascending=False)
```

df[['PARTY NAME', 'CLV']]: This picks out just the 'PARTY NAME' and 'CLV' columns from the data.

drop_duplicates(): This removes any repeated customer names, so each customer appears only once in the list.

sort_values(by='CLV', ascending=False): This arranges the list so that customers with the highest CLV are at the top.

I have then made a bar chart of the top 10 Customers by CLV

The results and finding of this has been given in *'Results, Findings, and Interpretation of Results'* section

Correlation matrix:

Correlation Matrix:

- It calculates the relationship (correlation) between different numerical variables in your dataset. The values range from -1 to 1:
  - 1 means a perfect positive correlation (as one variable increases, the other also increases).
  - -1 means a perfect negative correlation (as one variable increases, the other decreases).
  - 0 means no correlation (no relationship between the variables).

Detailed Steps Correlation matrix:

Made a correlation matrix to better understand the correlation between all the features in my dataset. Then plotted a heatmap for the same. The details for which has been attached in the *'Results, Findings, and Interpretation of Results'* section

Detailed Steps for Inventory Management analysis:

First, I performed preprocessing on the DATE column by formatting the dates according to my requirements and removing any rows with invalid dates. Then, I identified the months with the highest sales by sorting the data in descending order. Using this information, I created a Bar chart to visualize which months contributed most to the overall sales. This insight allows me to strategically focus on stocking more inventory during the peak months, ensuring the shop doesn't run out of stock during those high-demand periods. Additionally, it provides guidance on which months require less inventory, optimizing stock levels during lower sales periods.

The results for this has been attached in the *'Results, Findings, and Interpretation of Results'* section

Churn analysis:

I decided to churn to do churn analysis here better understand the following:

1. **Retain Customers**: Identify why customers leave and take action to keep them.
2. **Boost Revenue**: Reducing churn is cheaper than acquiring new customers
3. **Improve Services**: Understand customer needs to enhance offerings.

Detailed Steps for Churn analysis:

At first I filtered the data by date range, start date = '2021-04-01' and end date = '2022-03-31'. Then I grouped the data by the 'PARTY NAME' column. Then I counted the number of bills (i.e., transactions) per party and summed the total bill amount for each party. Then reseted the index to convert the grouped data back into a regular DataFrame.

Then I filtered out the data to include only transactions that occurred in the last quarter (December 1, 2021, to March 31, 2022).

Identifies customers who did not make any purchases during the last quarter by finding the difference between all customers and those who made purchases in the last quarter.

After that applied a function to each 'PARTY NAME'. If a party is in the `churned_parties` set, they are marked as 'Yes' for churn; otherwise, they are marked as 'No'.

Then i visualized the analysis with the help of a bar chart which can be observed in the *'Results, Findings, and Interpretation of Results'* section

3. Results, Findings, and Interpretation of Results

RFM analysis chart categorizes customers based on their purchasing behavior, which can help tailor strategies to improve sales at DEY MEDICO shop. Here's an elaboration on the findings from the chart:

1. Loyal Customers:
    - Largest segment, indicating a strong base of repeat customers.
    - Focus on retention strategies, loyalty programs, and personalized communications to maintain and enhance their engagement.
2. Potential Loyalists:
    - Substantial group, these customers are close to becoming loyal.
    - Convert them into loyal customers through targeted promotions, upselling, and excellent customer service.
3. Lost:
    - Significant number of customers who haven't made a purchase recently.
    - Initiatives such as reactivation campaigns, personalized offers, and feedback collection could win them back.
4. Champions:
    - Smaller but vital group who spend the most and are highly engaged.
    - Keep these high-value customers satisfied with exclusive deals, early access to new products, and invitations to special events.
5. At Risk:
    - These customers are likely to churn.
    - Prioritize outreach with incentives and check for underlying issues that could be addressed to retain them.
6. New Customers:
    - Smallest segment, indicating room for growth in acquiring new customers.
    - Implement acquisition strategies like introductory offers, social media marketing, and referral programs to expand this segment.

Customer Segmentation Distribution
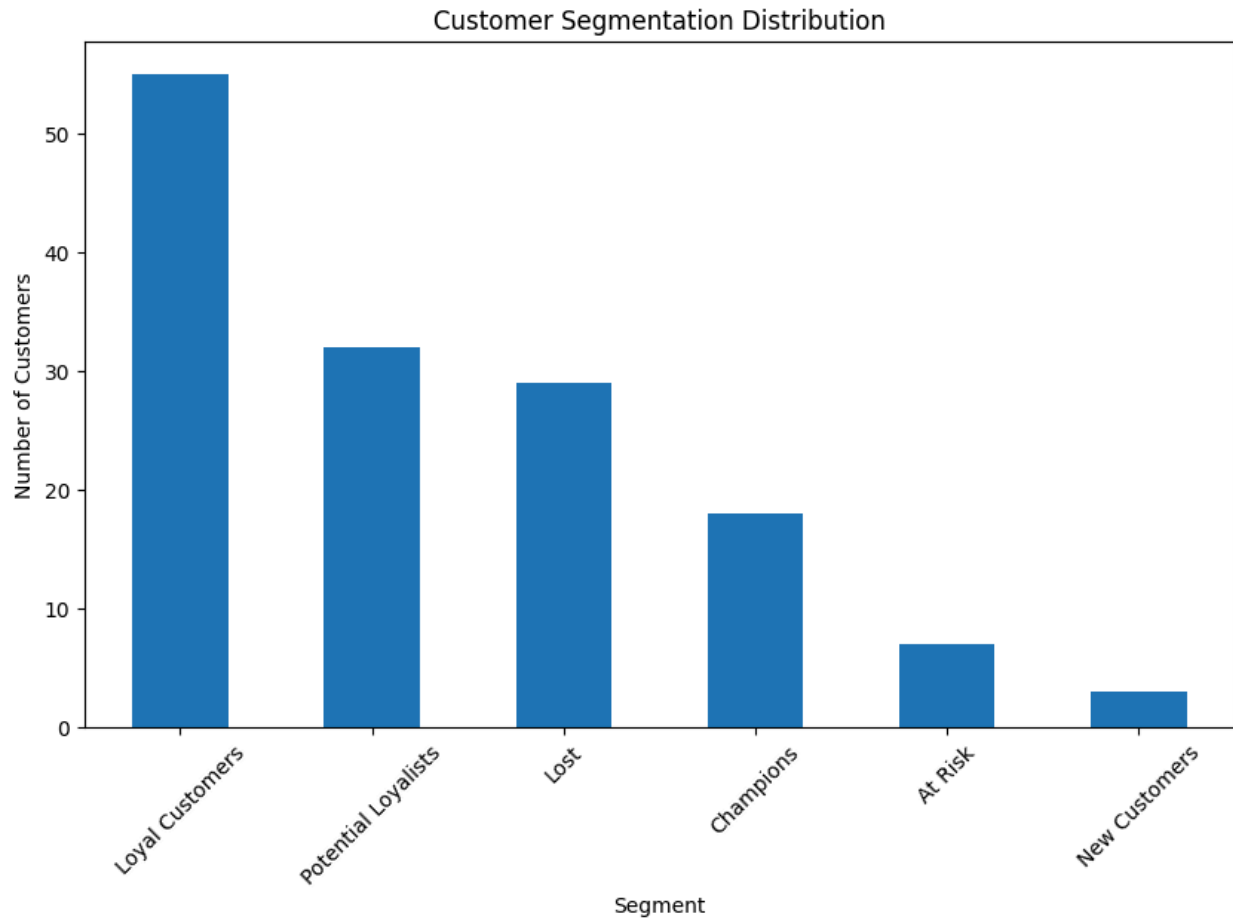
RFM correlation heatmap

Additionally I made a RFM correlation heatmap which provides a visual representation of the relationship between the three RFM dimensions: Recency, Frequency, and Monetary value.

Recency:

- Recency has a weak negative correlation with Frequency (-0.3) and Monetary (-0.37). This suggests that customers who purchased recently do not necessarily buy more frequently or spend more.

Frequency:

- Frequency has a strong positive correlation with Monetary (0.92). This indicates that customers who purchase more frequently tend to spend more overall.

Monetary:

- As expected, Monetary correlates strongly with Frequency, reinforcing the idea that higher purchasing frequency leads to higher spending.
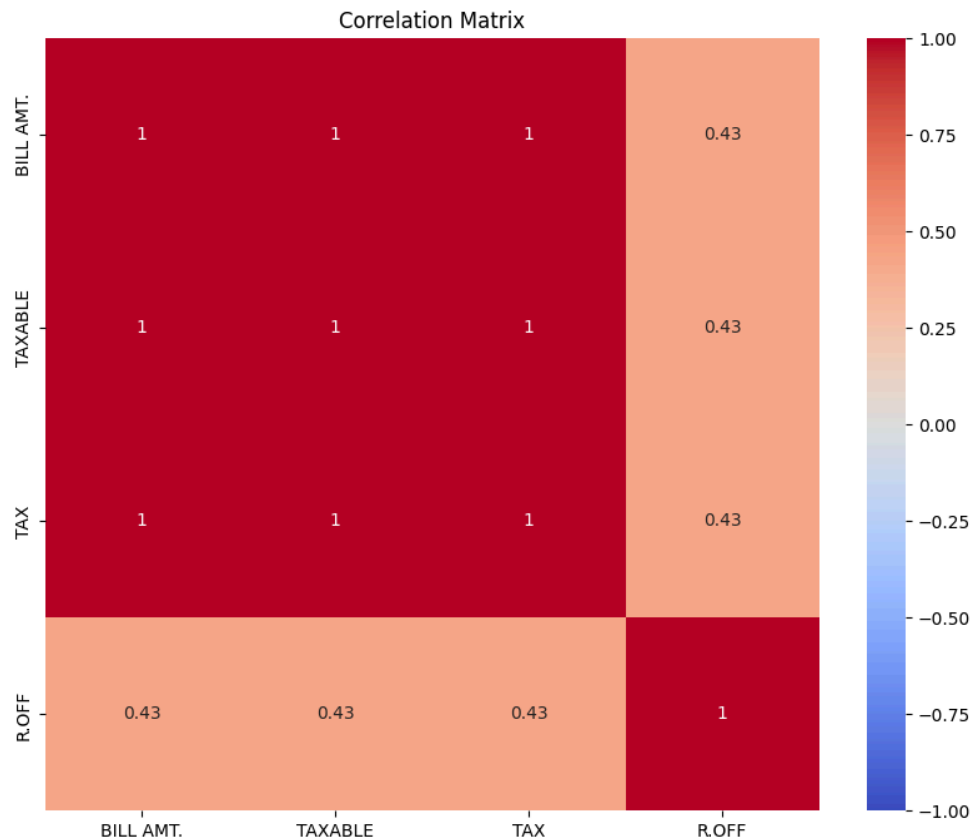

RFM Correlation Heatmap
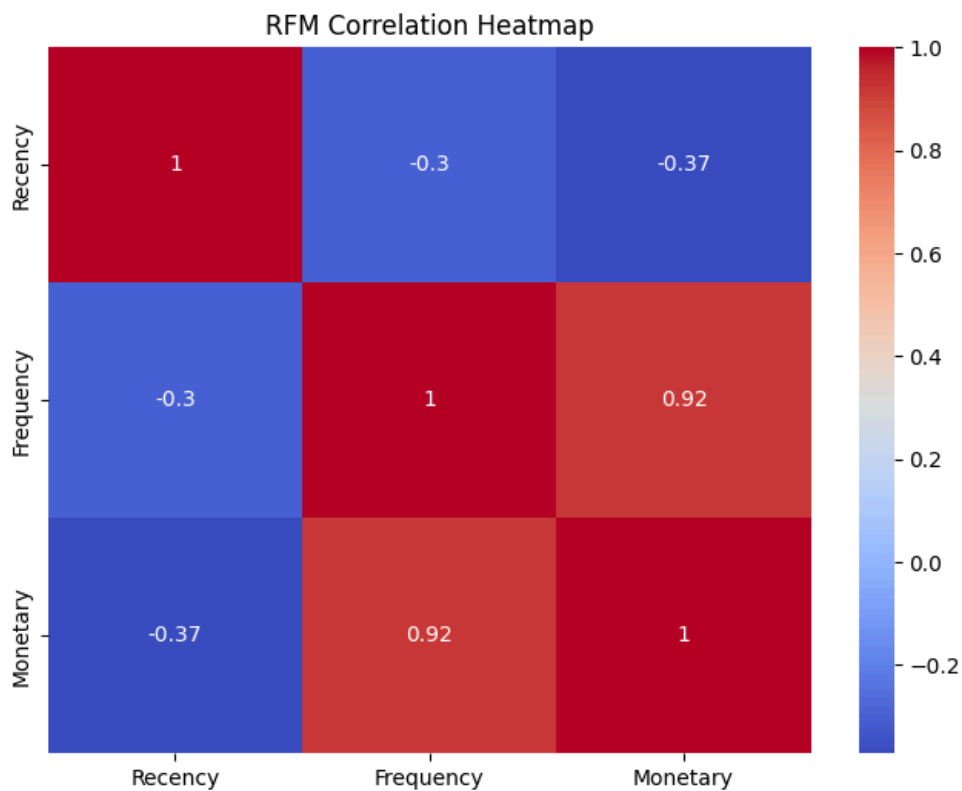
Customer Lifetime Value(CLV) findings:

Top 10 Customers by CLV
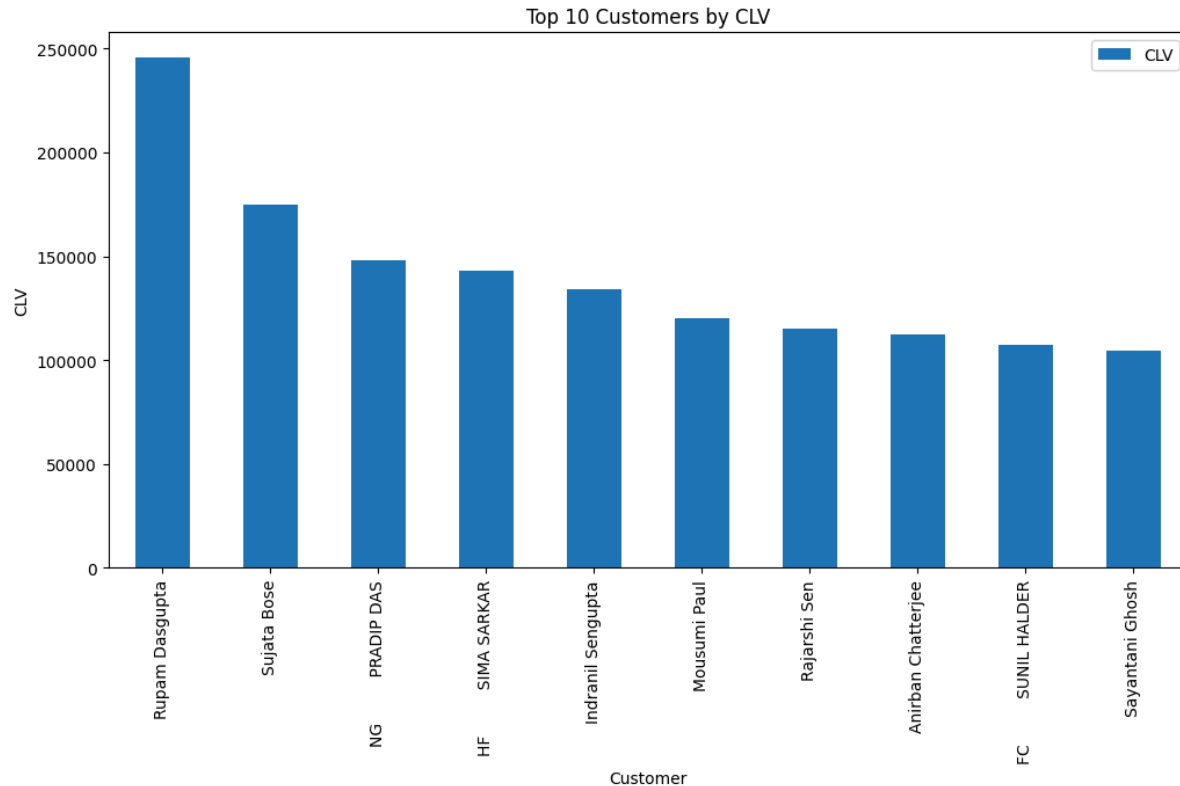
"Top 10 Customers by CLV" (Customer Lifetime Value) chart:

1. **Top Contributors**: Rupam Dasgupta and Sujata Bose are the most valuable customers, with Rupam Dasgupta leading by a significant margin.
2. **Revenue Concentration**: A small number of customers contribute a substantial portion of the revenue, indicating the importance of these top 10 customers.
3. **Strategic Opportunities**: Targeted marketing and retention strategies should focus on these high-CLV customers to further boost their value and loyalty.
4. **Risk Management**: The business's heavy reliance on a few key customers highlights the need for strong retention efforts to mitigate potential risks if these customers churn.

Correlation matrix:

The heatmap provides insights into the relationships between the variables BILL AMT., TAXABLE, TAX, and R.OFF in your dataset. Here's what can be inferred:

1. Perfect Positive Correlation (1.0):
   - BILL AMT., TAXABLE, and TAX all have a perfect positive correlation of 1 with each other.
   - This indicates that these three variables move together perfectly. When BILL AMT. increases, both TAXABLE and TAX also increase in a perfectly predictable manner.
   - This is expected because TAXABLE and TAX are likely calculated based on the BILL AMT., leading to this perfect correlation.
2. Moderate Positive Correlation (0.43):
   - The variable R.OFF has a moderate positive correlation of 0.43 with BILL AMT., TAXABLE, and TAX.

- This suggests that while R.OFF does tend to increase when the other three variables increase, the relationship is not as strong or direct. There may be other factors influencing R.OFF that aren't captured by just BILL AMT., TAXABLE, and TAX.

Interpretation:
- The strong correlations between BILL AMT., TAXABLE, and TAX indicate that these variables are highly related, possibly because they are derived from or closely tied to each other.
- The moderate correlation between R.OFF and the other variables suggests that R.OFF is somewhat related to the bill amounts and taxes but not as directly as the other three variables are to each other.

Findings for Inventory management:

Top Performing Months:
- July and June have the highest sales, indicating peak demand during these months. These months should be prioritized for stocking more inventory to meet the high demand and avoid stockouts.

- April and May also show strong sales performance, suggesting that they too are critical months for maintaining a robust inventory.

Moderately Performing Months:
- March, January, November, and December show moderate sales levels. While these months do not reach the peak sales seen in July and June, they still require adequate stock to meet customer demand.

Low Performing Months:

● October, February, and especially August show lower sales figures. These months could be considered off-peak, and therefore, you might need to stock less inventory compared to the busier months.

Sales by Peak Month
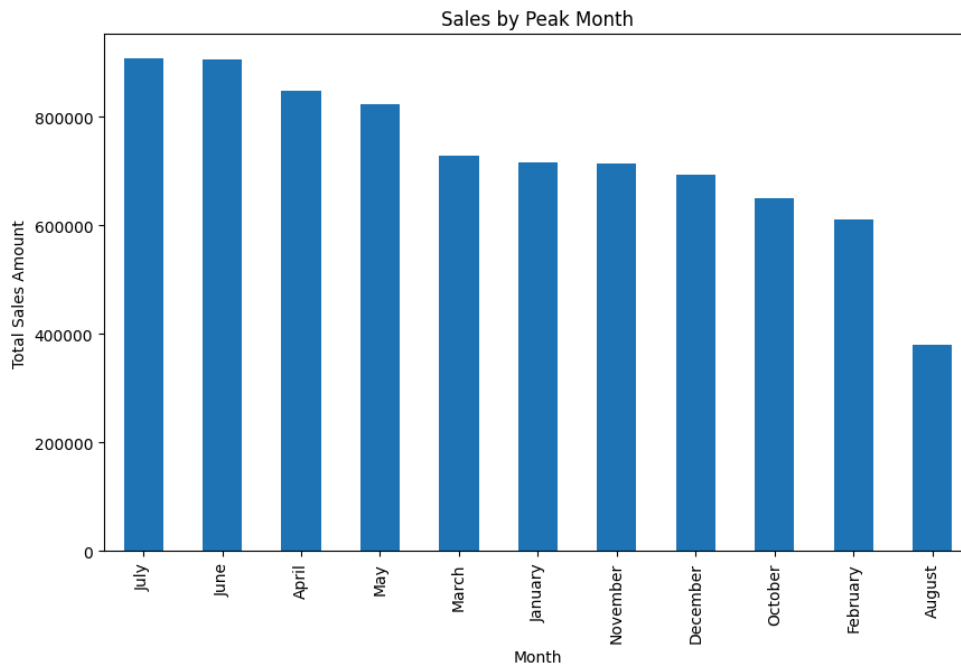
Churn Analysis findings:

Customer Churn Analysis
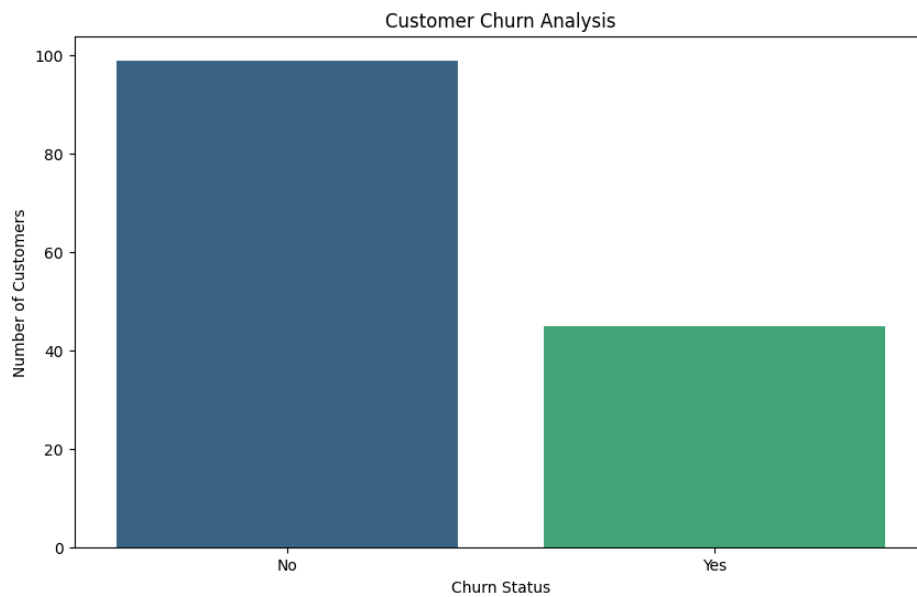
This bar chart represents the results of your customer churn analysis. Here's what it indicates:

- Churn Status (X-axis): Divides customers into two categories: those who did not churn ("No") and those who did ("Yes").
- Number of Customers (Y-axis): Shows the count of customers in each category.

Interpretation:

- "No" Bar: The taller bar indicates that most customers (around 100) did not churn, meaning they continued to make purchases.
- "Yes" Bar: The shorter bar shows that a smaller number of customers (around 40) churned, meaning they stopped making purchases during the last quarter.

4. Interpretation of Results and Recommendation

Key Takeaways:

- **Churn Analysis**: The low churn rate indicates that most customers are satisfied and continue to make purchases. However, the presence of churned customers suggests that some are disengaging, potentially due to unmet needs or attractive offers from competitors. **Recommendation**: Implement a targeted retention program that identifies at-risk customers early and engages them with personalized offers, loyalty rewards, or exclusive services. Regular follow-ups with customers who haven't purchased recently can help in reactivating them.
- **RFM Analysis**: The segmentation of customers into categories like "Loyal Customers," "Champions," and "At Risk" provides a clear roadmap for DEY MEDICO's customer engagement strategies. **Recommendation**:
  - For **Loyal Customers** and **Champions**, continue with high-touch strategies such as exclusive deals, personalized communications, and VIP programs to maintain their loyalty.
  - For **Potential Loyalists**, offer incentives for repeat purchases, such as discounts on their next purchase or a points-based reward system.

- For the **At Risk** and **Lost** customers, implement re-engagement campaigns with targeted marketing, perhaps offering a win-back discount or special promotion to entice them back.
- **Customer Lifetime Value (CLV) Analysis**: The CLV analysis highlights that a small group of customers contributes significantly to the revenue. **Recommendation**: Focus on maintaining strong relationships with these top customers through personalized services, special privileges, and direct communication to ensure they continue to feel valued and remain loyal. Additionally, explore ways to convert more customers into high-CLV customers by analyzing their behavior and offering tailored promotions.
- **Inventory Management**: The analysis of peak and low sales months offers insights into inventory optimization.

  Recommendation:

    - During high-demand months (June, July), ensure that stock levels are high enough to meet customer demand without risking stockouts.
    - In off-peak months, reduce inventory levels to avoid overstocking and carrying costs. Consider running promotions or bundling offers during these months to boost sales.
- **Correlation Analysis**: The correlation between frequency and monetary value suggests that encouraging more frequent purchases can drive higher overall revenue.

  Recommendation: Implement strategies to increase purchase frequency, such as loyalty programs where frequent purchases are rewarded, or time-limited promotions that create urgency.

- **Competitive Landscape**: Given the presence of competitors like Apollo Pharmacy, DEY MEDICO must remain vigilant about pricing and service quality.

  Recommendation: Conduct regular market analysis to ensure competitive pricing, and differentiate the store through superior customer service, personalized experiences, and unique product offerings.

Strategic Recommendations:

1. **Implement a Tiered Loyalty Program**: Introduce a tiered loyalty program where customers earn points for purchases that can be redeemed for discounts or exclusive products. This will encourage repeat purchases and increase customer engagement.
2. **Leverage Data for Personalized Marketing**: Use the RFM and CLV data to create personalized marketing campaigns that target different customer segments. Tailored emails, SMS offers, and social media ads can drive higher engagement and sales.
3. **Expand Digital Presence**: In an increasingly digital world, DEY MEDICO could benefit from enhancing its online presence. Whether through social media, a website, or an app, providing customers with easy access to information and online shopping options can boost sales and attract new customers.
4. **Monitor and Adjust**: Continuously monitor the impact of implemented strategies and adjust based on performance. Regularly revisit the analysis to ensure that the strategies remain aligned with changing customer behaviors and market conditions.

By adopting these strategies, DEY MEDICO can not only enhance its current operations but also secure a stronger position in the market, ensuring long-term growth and profitability.

The dataset which I used for my analysis can be found below:DATASET

All my work for the analysis can be found in the following colab link:COLAB