A Quantitative Analysis to Evaluate and Enhance Operational Efficiency and Profitability in a Pharmaceutical Retail Store

Mid-Term Submission for Business Data Management Capstone Project

Submitted by

Name: Shreya Shambhavi

Roll Number: 21F1001507

IIT Madras BS Degree Program,

Indian Institute of Technology Madras, Chennai

Tamil Nadu, India, 600036

Table of Contents

1. Executive Summary

This capstone project focuses on “Singh Medico”, a Pharmaceutical Retail Store located at Lalgutwa in Ranchi, Jharkhand. The store established in 2018 is owned by Mr. Rajeev Ranjan Singh. This B2C business primarily deals with medicinal drugs and other related FMCG products. Situated on the outskirts of Ranchi, Singh Medico serves a substantial geographical area due to the unavailability of another medical store nearby. Moreover, the location is witnessing many housing projects consequently adding on to the whole demand.

To help the business expand and in turn enhance its profitability, various analytical techniques shall be used to draw out significant inferences from the business data. The dataset consists of sales, purchase and stock data. The data collection process involved various visits to the store, increasing the overall interaction and understanding. The process of data cleaning involved handling missing categorical data points, structuring the dataset for further analysis steps. Performing Exploratory Data Analysis on the dataset helped understand the data and the involved attributes in a better way. Descriptive statistics help draw summary out of the features (column labels) involved in the dataset. Methods involving various Python libraries and packages shall be significant to examine the trends in sales and purchases. Detecting correlation or a pattern amongst the attributes will help us asses the demand and ways to handle involved shortcomings. Analysis the products that are adding the most to the sales would help detect ways to enhance the overall profitability of the business and a better inventory management.

2. Proof of Originality of Data

The handwritten letter by the owner along with pictures and video has been attached as the proof of originality of the data used for analysis purposes in the project.

For the google drive folder, kindly **Click Here**

3. Metadata and Descriptive Statistics

Dataset: Click Here

The dataset consists of the ***Sales and Purchase Data*** of the store for a time period of ***3 months***, ranging from ***01st October, 2023 – 31st December, 2023***. This is accompanied by the ***Stock Data*** of the store as on ***31st January, 2024***.

A. Sales Data

The shape of the dataset (total number of rows and columns respectively) is: **(5020, 14)**

Following are the features (column labels) for the sales data:

S. No.	Data Feature	Datatype	Comments
0	Serial Number	Categorical	-
1	Bill Number	Categorical	Bill number of a particular sold product. The data consists of a total of **289** bills.
2	Order Month	Categorical	Month of a particular sold product. It ranges from **Oct – Dec**.
3	Product Name	Categorical	Name of the particular medicine sold. The data consists of a total of **5020** items.

4	Manufacturing Company	Categorical	Name of the manufacturing company of a particular medicine item.
5	Pack	Categorical	The packaging a particular medicine is available in the market. For example, 1 X 10 TABS.
6	Expiration Date	Categorical	Expiration date of a particular medicine sold.
7	MRP	Numerical	Maximum retail price of a particular product item.
8	Quantity	Numerical	Total quantity of a particular medicine sold to the customer.
9	Discount Percent	Numerical	Discount percentage given on a particular medicine.
10	Discount Amount	Numerical	Discount amount given on a particular medicine.
11	Total Discount	Numerical	Total discount given on a particular bill.
12	Grand Total	Numerical	Grand total of the amount paid by a customer on a particular bill.
13	Payment Mode	Categorical	Payment mode used by the customer. It's a binomial category consisting of CASH and GENERAL LEDGER.

B. Purchase Data

The shape of the dataset (total number of rows and columns respectively) is: **(905, 11)**

Following are the features (column labels) for the purchase data:

S. No.	Data Feature	Datatype	Comments
0	Serial Number	Categorical	-
1	Purchase Month	Categorical	Month of a particular purchased product. It ranges from **Oct – Dec**.
2	Product Name	Categorical	Name of the particular medicine purchased.
3	Manufacturing Company	Categorical	Name of the manufacturing company of a particular medicine item.
4	Pack	Categorical	Packaging a particular medicine is available in the market.
5	Expiration Date	Categorical	Expiration date of a particular medicine bought.
6	Quantity	Numerical	Total quantity of a particular medicine bought from suppliers.
7	MRP	Numerical	Maximum retail price of a particular product item.
8	Rate	Numerical	Price at which a particular product was bought from suppliers.

9	Discount Percent	Numerical	Discount percent given on a particular item.
10	Grand Total	Numerical	Grand total of a particular item bought i.e. quantity x rate.

C. Stock Data

The shape of the dataset (total number of rows and columns respectively) is: **(5719, 18)**

Following are the features (column labels) for the stock data:

S. No.	Data Feature	Datatype	Comments
0	Serial Number	Categorical	-
1	Item ID	Categorical	Item ID of an item.
2	Product Name	Categorical	Name of the medicinal item.
3	Manufacturing Company	Categorical	Name of the manufacturing company.
4	Item Code	Categorical	Item code of a particular product.
5	HSN Code	Categorical	The Harmonized System of Nomenclature Code of an Item.
6	HSN Name	Categorical	HSN name of an item.
7	Local Tax	Categorical	Depicts whether local tax is applicable on the item.

8	Central Tax	Categorical	Depicts whether central tax is applicable on the item.
9	SGST	Numerical	State Goods and Services Tax
10	CGST	Numerical	Central Goods and Services Tax
11	IGST	Numerical	Integrated Goods and Services Tax
12	P Rate	Numerical	Price per piece/ tab/ capsule.
13	MRP	Numerical	The maximum retail price of a particular product item.
14	Balance	Numerical	The inventory balance of a particular product.
15	Diff Tax	Numerical	Difference of taxes
16	Category	Categorical	Category of a particular item
17	Salt	Categorical	Salt composition of a particular medicine

Now that we are aware of our data, let's explore the same to draw some basic meaningful inferences that would help us understand it in a better way. We would be using measures of central tendency and dispersion on the numerical data columns to detect trends and any anomaly present in the data. This step would enhance our understanding of the data points and in turn will help choose the better method and attribute for further analysis. We would be going through sales, purchase and stock data one after another, highlighting any pattern and anomaly detected such as outliers.

1. Measures of Central Tendency

A. Sales Data

S. No.	Column Label	Mean	Median	Mode
1	MRP	115.36635	90.00	65.00
2	Quantity	1.19012	1.00	1.00
3	Discount Percent	6.17978	5.00	0.00
4	Discount Amount	8.63515	1.76	0.00
5	Total Discount	301.75362	245.30	349.24
6	Grand Total	3522.97310	3622.00	5677.00

B. Purchase Data

S. No.	Column Label	Mean	Median	Mode
1	Quantity	7.38784	3.00	2.00
2	MRP	180.73725	130.00	90.00
3	Rate	121.74917	87.43	**0.01, 58.57**
4	Discount Percent	3.21039	3.00	0.00
5	Grand Total	462.94869	281.42	**270.00, 600.00**

C. Stock Data

S. No.	Column Label	Mean	Median	Mode
1	Balance	18.98898	3.00	0.00
2	MRP	204.87574	131.00	150.00
3	P Rate	131.03850	83.50	100.00

2. Measures of Dispersion

A. Sales Data

S. No.	Column Label	Min	Max	25%	50%	75%	St Dev
1	MRP	1.00	3400.00	45.00	90.00	132.68	127.36
2	Quantity	0.05	**300.00**	0.40	1.00	1.00	5.23
3	Discount Percent	0.00	**100.00**	0.00	5.00	10.00	8.94
4	Discount Amount	0.00	1700.00	0.00	1.76	9.9	41.54
5	Total Discount	0.00	2173.33	166.36	245.30	331.98	335.54
6	Grand Total	30.00	6931.00	2729.00	3622.00	4522.00	1543.65

B. Purchase Data

S. No.	Column Label	Min	Max	25%	50%	75%	St Dev

1	Quantity	1.00	240.00	2.00	3.00	6.00	18.04
2	MRP	1.00	**1775.00**	75.67	130.00	211.95	184.74
3	Rate	0.01	**1170.71**	42.80	87.43	148.07	128.61
4	Discount Percent	0.00	41.25	0.00	3.00	5.00	3.11
5	Grand Total	0.00	6336.00	154.28	281.42	499.29	616.47

C. Stock Data

S. No.	Column Label	Min	Max	25%	50%	75%	St Dev
1	Balance	**-90.00**	1857.00	1.00	3.00	20.00	50.34
2	MRP	0.00	**10000.0**	75.00	131.00	228.48	315.27
3	P Rate	0.00	**5500.00**	39.42	83.50	150.57	183.84

D. Detailed Explanation of Analysis Process/Method

1. Data Collection

The dataset was collected over a duration of a few weeks through several visits to the store. Mr. Singh uses a software for storing the sales, purchase and stock data. Regrettably, the data files were not in a structured exportable form which can be straight forwardly downloaded and used by simply loading the same in an excel worksheet or a python notebook. Pattern extraction using the command line was also quite an impossible task. Hence, it was required to manually go through each and every bill of sales and purchases involved for the duration of 3 months on the software itself to build up the required excel sheet. Consequently, the step of data collection took more that the earlier intended time.

2. Data Cleaning

Since sales and purchase data were gone through manually during the process of data collection, they lacked the presence of any missing data. On the other hand, stock data was obtained as a .csv file generated by the store software. It involved various missing data points for 4 categorical columns, which were, Item Code, Product Name, HSN Name and Salt. The missing datapoints where handled by filling them up with the *mode value* of that particular column. Further on, it was ensured that all datatypes are correct for each column (for example, MRP, Balance, Grand Total etc were stored as floats or integers.) This was followed by taking care of inconsistencies involved in the dataset such as duplicates.

3. Exploratory Data Analysis

EDA was performed on the sales, purchases and stock data. This step was primarily done using python libraries on a google collab notebook. It involved the calculation of basic statistics like mean, median, mode, min, max, 25%, 50%, 75%, standard deviation in order to understand the central tendency and dispersion involved among the data points for a particular numerical feature of the dataset. Consequently, this step helped point out the presence of outliers present across the dataset that are further required to be taken care of for further analysis.

E. Results and Findings

1. Important conclusions drawn out after performing EDA on the datasets include: a) 'Rate' and 'Grand Total' columns present in the purchase data are bimodal in nature. b) 'Quantity', 'Discount Percent' in sales data; 'MRP', 'Rate' in the purchase data and 'Balance', 'MRP' and 'P Rate' have min/max value that is quite away when compared to 25% or 75% respectively. This concludes the presence of outliers in the involved attributes.

2. When plotting total purchases in terms of rate of the item bought in accordance to purchase month, we observed a high purchase in the month of October which further on dipped significantly in November and December.

3. When plotting total sales in terms of MRP of the item sold grouped by order month, we observed an increase from October to November which was closely followed by a dip in December. It is required to be further analysed in order to under the shortcomings.

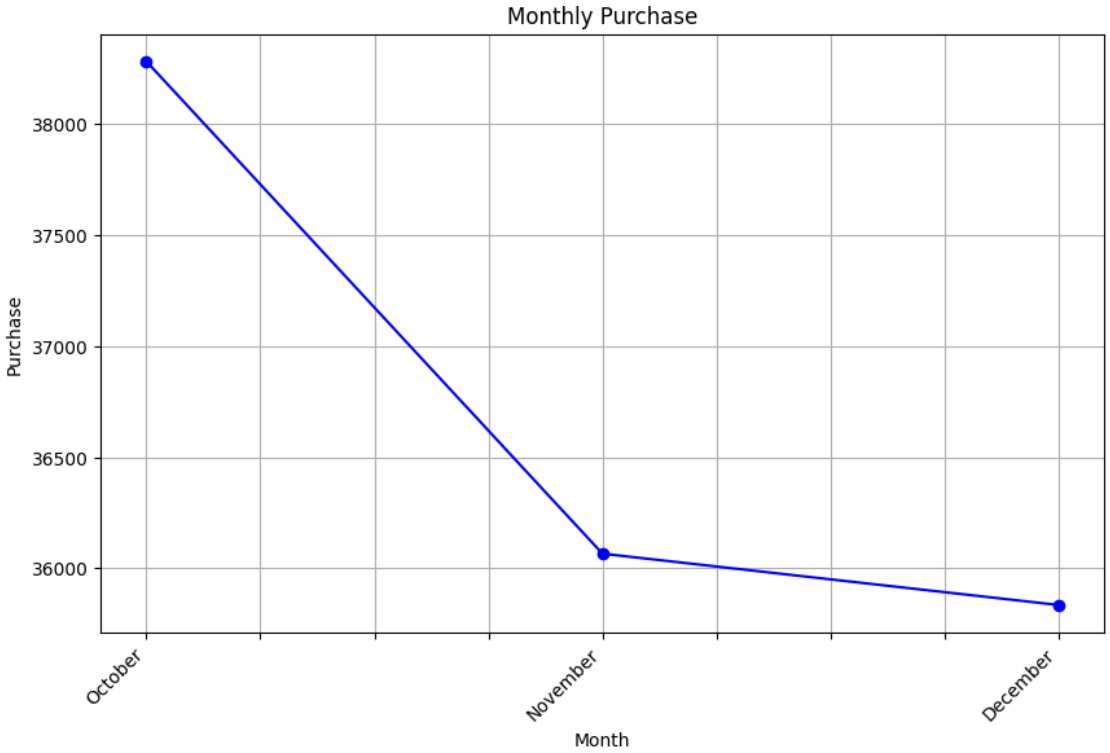

Figure 1: Monthly purchase obtained by plotting in term of rate of items bought.

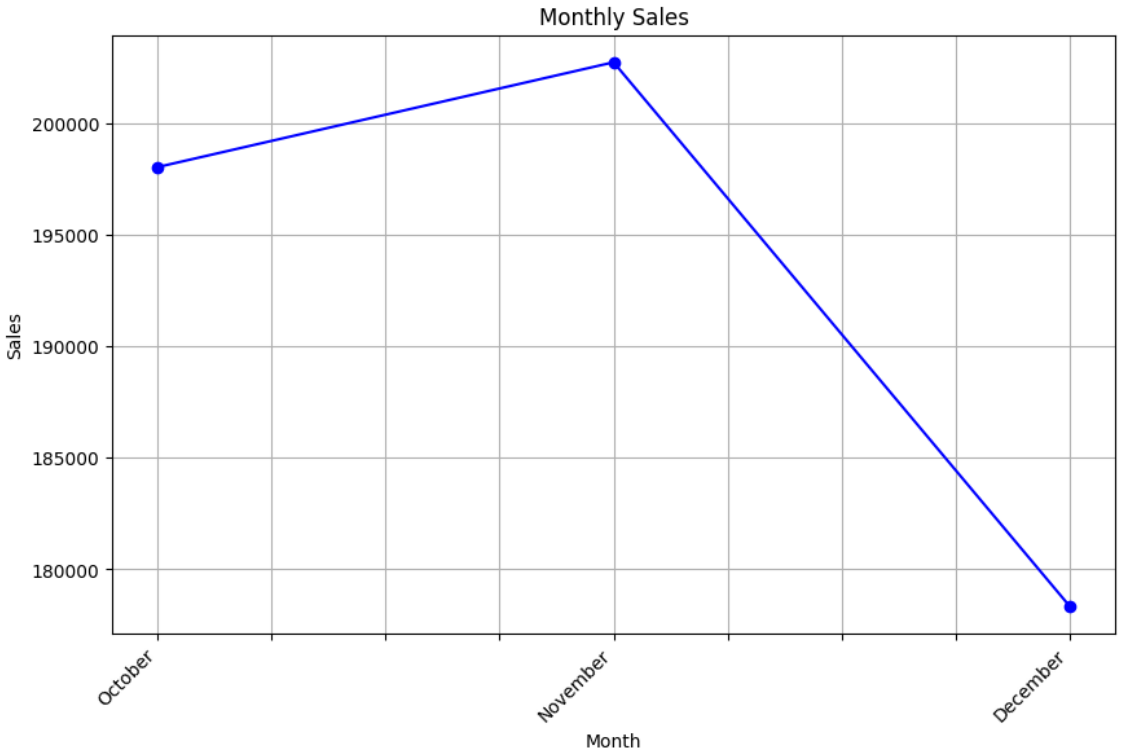

Figure 2: Monthly sales obtained by plotting in term of MRP of items sold

Note: The analysis is an ongoing process right now. Further concrete inferences will be discussed in the Final Report.