LEVERAGING DATA FOR EFFECTIVE PRICE FORECASTING AND CREDIT MANAGEMENT FOR VEGETABLE DISTRIBUTOR

BDM CAPSTONE PROJECT MID-TERM SUBMISSION

ADITYA JAISWAL
22F3002960

MAY 2024 TERM

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS, CHENNAI
TAMIL NADU, INDIA, 600036

(BS) DEGREE IN DATA SCIENCE AND APPLICATIONS

INDEX

1. Executive Summary

The midterm report focuses on leveraging data analysis for effective price forecasting and credit management for Kanhaiyalal D Jaiswal, a wholesaler of onions and potatoes. The business faces key challenges, including managing seasonal price fluctuations, inventory control, and an unstructured credit allocation process. The report proposes a data-driven approach using linear regression models to forecast price trends and improve inventory management. The project aims to provide clear insights into market trends, helping the business make informed decisions to improve financial planning and customer service.

The dataset comprises sales data from June 2023 to June 2024, containing 594 records. It includes fields like order date, product type, quantity (in bags and kgs), cost per 10 kgs, and total bill. Descriptive statistics reveal significant variability in order sizes, with the average quantity being 14.26 bags or 745.59 kgs. Prices also vary widely, with the cost per 10 kgs ranging from ₹80 to ₹580. These statistics provide a foundation for understanding customer behavior, pricing trends, and total sales volume..

The report also describes the step-by-step process of data collection, cleaning, and analysis. The initial challenge was converting offline data from accounting books into a digital format. The report highlights the use of spreadsheets for data aggregation and analysis, alongside time-series analysis for financial forecasting. Python libraries like Pandas and Matplotlib were instrumental in conducting descriptive statistical analysis and data visualization. Engaging with the business owner provided qualitative insights, enabling a practical approach to problem-solving.

The analysis provided valuable insights into sales composition, price trends, and seasonality. A pie chart of onion versus potato sales showed that onions are in significantly higher demand. Monthly sales fluctuated, with November registering the highest sales of ₹14 lakh, and July experiencing the lowest at ₹4 lakh. The price trend analysis revealed that onion prices were more volatile than potato prices, peaking at ₹500 per 10 kg in October. The report emphasizes the importance of understanding these trends for effective decision-making in inventory management, pricing, and credit allocation.

2. Proof of Originality

To establish the authenticity of the data, the supporting evidence as listed below:

1. **Letter from the Organization:** Access to the letter can be obtained through the G-Drive link: [https://drive.google.com/drive/folders/1GGmaZxC9zegeSzvEr42eIL1e2qSCXRiO?usp=drive_link]
2. **Interview with Owner :** They can be accessed through the G-Drive link: [https://drive.google.com/drive/folders/1GKDpSJYUPtf0h2sagVn2oxCooHPsXtrY?usp=drive_link]
3. **Images of Organization:** They can be accessed through the G-Drive link : [https://drive.google.com/drive/folders/1GBiV_NC3BFJ8SdPGzhbzoKCn_muM1bZ4?usp=drive_link]

3. Metadata and Descriptive Statistics

3.1 METADATA:

To effectively manage and understand the data sheet, the following metadata can be defined:

- **Dataset Overview**

  - **Title:** Kanhaiyal D Jaiswal Sales Data (June 2023 - June 2024)
  - **Description:** This dataset contains detailed information about orders for onions and potatoes, including dates, order types, quantities, prices, and total bills over a one-year period.
  - **Number of Records:** 594 rows - 6 columns
  - **Link:** https://docs.google.com/spreadsheets/d/112z_ZcB0QYhrgKyUDQM7qX-qTgFZAqB14YJZ47FUEbg/edit?usp=sharing

- **Columns and Their Metadata**

  Order Date

  - **Description:** The date on which the order was placed.
  - **Format:** DD-MM-YYYY
  - **Data Type:** Date
  - **Example:** 18/06/2024

  Category

  - **Description:** The type of product ordered, either onions or potatoes.
  - **Format:** Categorical (Onion, Potato)
  - **Data Type:** String
  - **Example:** Onion

  Quantity(In Bags)

  - **Description:** The quantity of product ordered, measured in no of bags(“goni” in local terms). Common unit to gauge amount of product ordered.
  - **Format:** Numeric
  - **Data Type:** Integer
  - **Example:** 25

Quantity(In Kgs)

- **Description:** The quantity of the product ordered, measured in kilograms (kgs).
- **Format:** Numeric
- **Data Type:** Integer
- **Unit:** Kilograms (kgs)
- **Example:** 836

Cost (Per 10 kgs)

- **Description:** The selling cost of the product. It is the price at which the product was sold.
- **Format:** Numeric
- **Data Type:** Integer
- **Unit:** Currency
- **Example:** 200

Total

- **Description:** The total bill amount for that particular order.
- **Format:** Numeric
- **Data Type:** Integer
- **Unit:** Currency
- **Example:** 51736

- **Additional Metadata**

Data Collection Period

- **Description:** The period during which the data was collected.
- **Start Date:** 01-06-2023
- **End Date:** 30-06-2024

Data Source

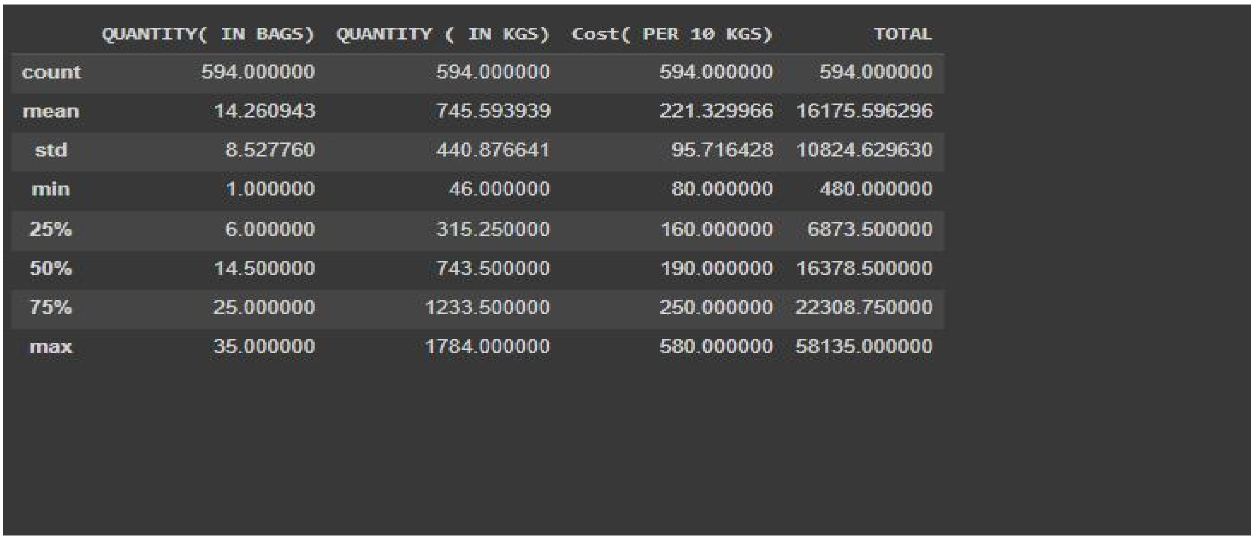
- **Description:** The origin of the data.
- **Source:** Sales records from Kanhaiyal D Jaiswal.

Data Quality

- **Description:** Information about the accuracy and completeness of the data.
- **Completeness:** 100%
- **Accuracy:** Assumed accurate as per business records.

3.2 DESCRIPTIVE STATISTICS:

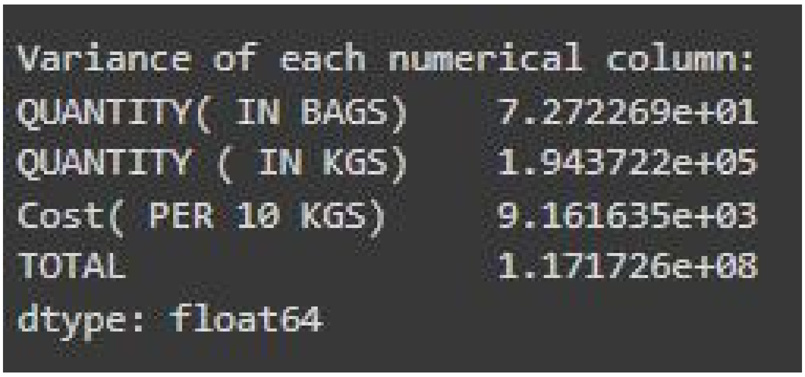
After data processing and cleansing, the following is a concise overview of the dataset using descriptive statistics.



	QUANTITY(IN BAGS)	QUANTITY (IN KGS)	Cost(PER 10 KGS)	TOTAL
count	594.000000	594.000000	594.000000	594.000000
mean	14.260943	745.593939	221.329966	16175.596296
std	8.527760	440.876641	95.716428	10824.629630
min	1.000000	46.000000	80.000000	480.000000
25%	6.000000	315.250000	160.000000	6873.500000
50%	14.500000	743.500000	190.000000	16378.500000
75%	25.000000	1233.500000	250.000000	22308.750000
max	35.000000	1784.000000	580.000000	58135.000000

(3.2.1) Descriptive Statistics

- **Quantity (in Bags and Kgs)**: The average order size is approximately 14.26 bags or 745.59 kgs. However, the orders exhibit high variability, with a standard deviation of 440.88 kgs. The smallest order is just 46 kgs, while the largest is 1,784 kgs, suggesting a wide range of customer purchasing behavior. The median order is 743.5 kgs, indicating that half the orders are above or below this quantity.
- **Cost per 10 Kgs**: The average cost per 10 kgs is ₹221.33, with prices ranging from ₹80 to ₹580. This reflects variability in product pricing, possibly due to differences in product type or quality. The majority of the orders (75th percentile) are priced at or below ₹250.
- **Total Cost:** The total cost of orders spans from ₹480 to ₹58,135, showing a large range in the size of transactions. The average order total is ₹16,175.59, with 50% of the orders costing at least ₹16,378. This wide range in total costs indicates diverse purchasing behaviors, possibly influenced by the varying order quantities and product types.
- **Spread of Data:** The interquartile range (IQR) for quantity in kgs (315.25 to 1,233.5) and for total cost (₹6,873.5 to ₹22,308.75) suggests a significant spread, highlighting diverse order sizes and corresponding costs.



```
Variance of each numerical column:
QUANTITY( IN BAGS)    7.272269e+01
QUANTITY ( IN KGS)    1.943722e+05
Cost( PER 10 KGS)     9.161635e+03
TOTAL                 1.171726e+08
dtype: float64
```

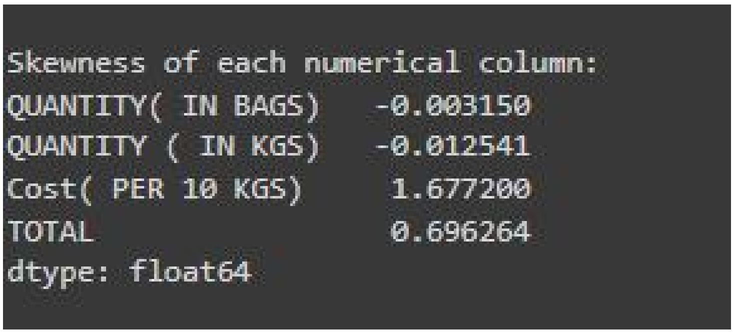
(3.2.2) Variance Statistics

- Quantity (in Bags): The variance of 72.73 suggests moderate variability in the number of bags ordered, indicating some consistency in the size of orders, though there is room for fluctuation.

- Quantity (in Kgs): The variance of 194,372.2 is quite large, showing substantial variability in the total weight of orders. This reflects that customers place a wide range of orders in terms of kgs.

- Cost per 10 Kgs: A variance of 9,161.64 indicates moderate variability in the price per 10 kgs. While prices do fluctuate, they don't vary as drastically as the order quantities.

- Total Cost: The variance of 117,172,600 reveals significant variability in the total cost of orders. This suggests a broad range of transaction sizes, likely influenced by differences in both order quantities and prices.



```
Skewness of each numerical column:
QUANTITY( IN BAGS)   -0.003150
QUANTITY ( IN KGS)   -0.012541
Cost( PER 10 KGS)     1.677200
TOTAL                 0.696264
dtype: float64
```

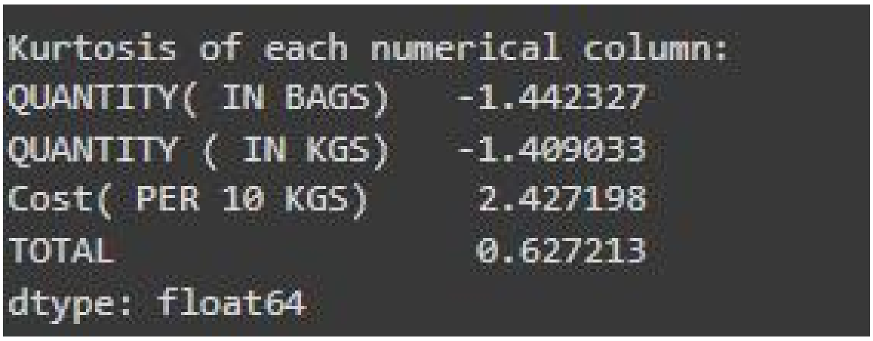
(3.2.3) Skewness Statistics

- Quantity (in Bags): The skewness of -0.003 suggests that the distribution of the number of bags ordered is almost perfectly symmetric, indicating a balanced spread of smaller and larger orders.

- Quantity (in Kgs): The skewness of -0.013 is similarly close to zero, showing a nearly symmetric distribution in the total weight of orders. There is no strong skew towards particularly large or small orders.

- Cost per 10 Kgs: The skewness of 1.68 indicates a positively skewed distribution, meaning that most prices are clustered at the lower end, but there are some notably higher prices pushing the distribution's tail to the right.

- Total Cost: The skewness of 0.70 shows a moderate positive skew in total order costs. While most order totals are lower, a few higher total costs stretch the tail on the right side.



```
Kurtosis of each numerical column:
QUANTITY( IN BAGS)   -1.442327
QUANTITY ( IN KGS)   -1.409033
Cost( PER 10 KGS)     2.427198
TOTAL                 0.627213
dtype: float64
```

(3.2.4) Kurtosis Statistics

- Quantity (in Bags): The kurtosis value of -1.44 indicates a platykurtic distribution, meaning the distribution has lighter tails and is flatter than a normal distribution. There are fewer extreme values in the order quantity (in bags) data.

- Quantity (in Kgs): The kurtosis of -1.41 also indicates a platykurtic distribution, suggesting that the order quantities in kilograms are similarly spread out with fewer outliers compared to a normal distribution.

- Cost per 10 Kgs: The kurtosis of 2.43 suggests a leptokurtic distribution, meaning the data has heavier tails than a normal distribution. This indicates the presence of more extreme price values, with prices spiking higher than the average more frequently.

- Total Cost: The kurtosis of 0.63 shows a mild leptokurtic tendency, implying a moderate level of outliers in the total cost data, although not as extreme as the cost per 10 kgs.

4. Analysis processes and methods

The process of data analysis encompasses defining the problem, data collection, organization, cleaning, transformation, applying analysis techniques, and drawing conclusions. The analysis process for the project also involves a combination of quantitative and qualitative methods, each chosen for their ability to address specific aspects of the business' challenges.

The first step of this journey was an arduous process of data acquisition, which proved to be the most challenging. Engaging with multiple businesses, and encountering a mix of rejections and hesitations, with some willing to share data verbally, but hesitant about sharing any raw data. Finally, a gracious business owner was willing to share the data, establishing one of the most important milestones in the journey of data analysis: data acquisition.

The next problem encountered was that the raw data was available only in offline format in the form of accounting books and laal khata books. There was no data that was maintained digitally either in the form of spreadsheets or any other form online. This made the data acquisition problem much more challenging as I had to manually input each data record from the accounting books to a spreadsheet . This took up a lot of time as the I had to enter a the data for a year from June 2023 to June 2024 but the benefit of this approach was since the data was manually entered by me there was hardly any need for data cleaning or data pre processing . The data entered was near 100% accurate and clean which saved up a lot of time that would have taken in cleaning the data.

Extensively utilized spreadsheets and their functions for various calculations essential to the analysis process. Spreadsheets provided a familiar and user-friendly interface for conducting complex calculations and aggregating data. Functions such as SUM, AVERAGE, etc were employed to streamline the process of summarizing and analyzing large datasets.

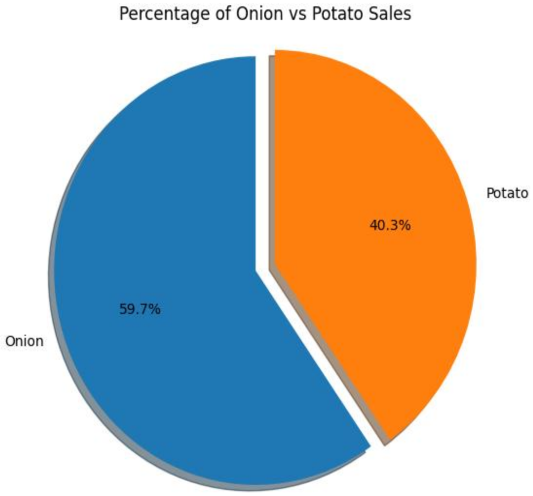
Time-Series Analysis: This method is particularly suitable for financial data, which is inherently time- dependent. By examining trends, patterns, and variations over time, we can gain insights into the Business' financial health and performance. This method stands out because it allows for understanding of trends based on historical data, which is crucial for forecasting the price of upcoming months and help making informed business decisions.

Python, along with libraries like Pandas , Numpy, was instrumental in conducting descriptive statistical analysis. Through Pandas, I computed measures of central tendency and variability, enabling us to understand the distribution of financial data points and identify any outliers or anomalies. Also libraries like Matplotlib and Seaborn were used which offer significant advantages in data visualization. They provide a powerful and flexible toolkit for creating a wide range of static, animated, and interactive plots with ease.

Conversations: Engaging with the business owner provides qualitative insights that are not captured by quantitative data alone. Understanding the owner's perspective on various problems is essential for tailoring recommendations that are practical and actionable

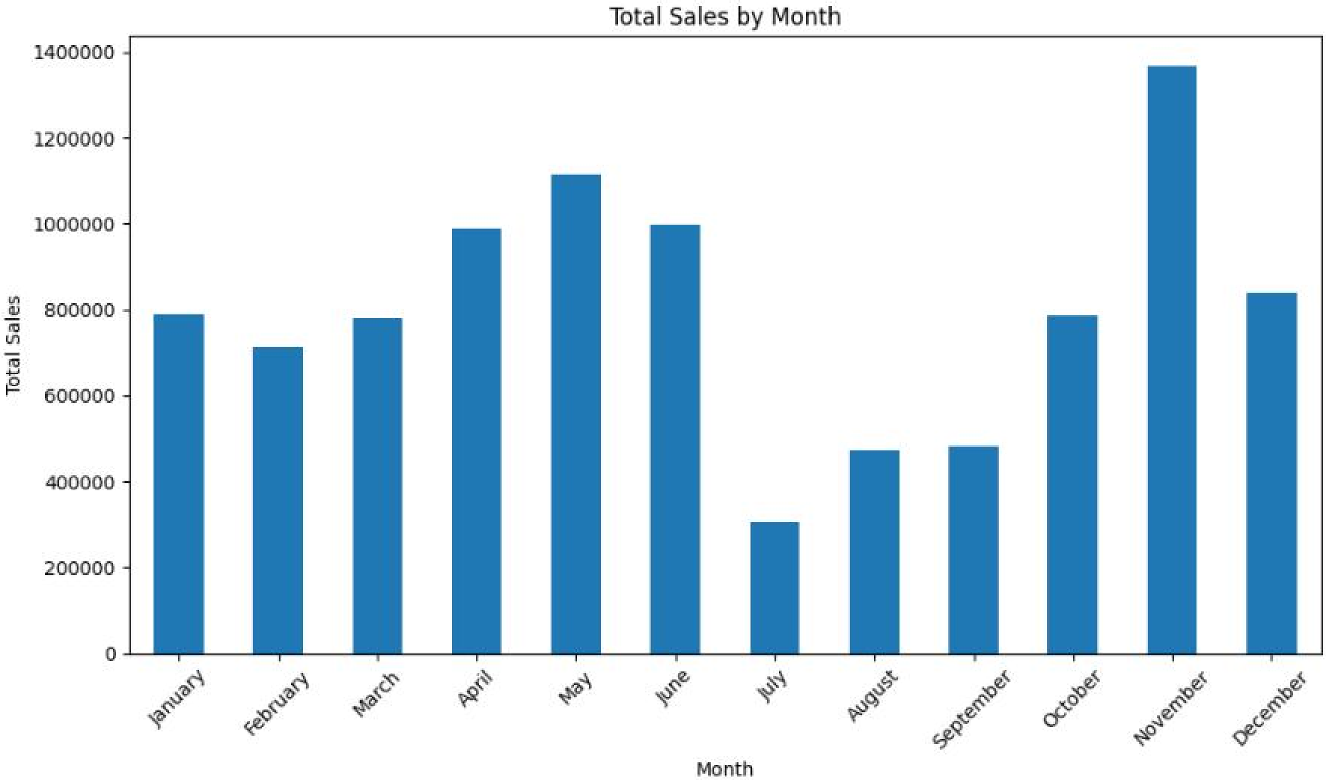
5. Results and Findings

Some insights gained from sales data:



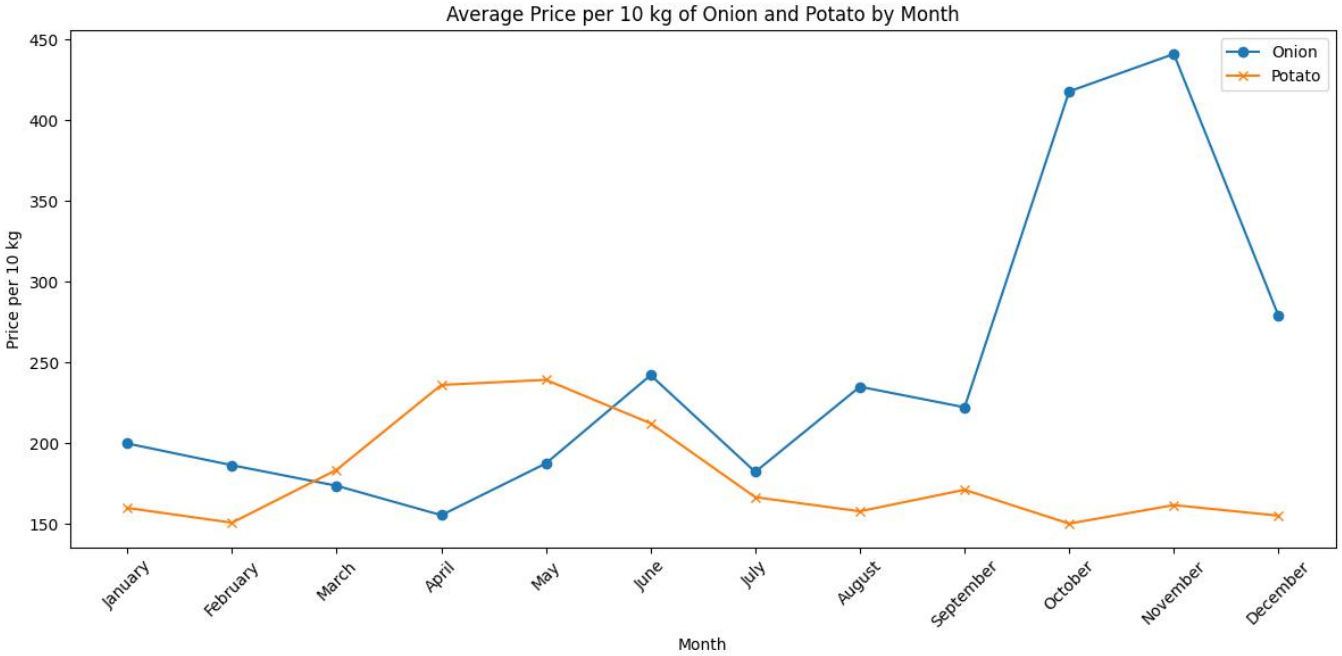
(5.1) Percentage Sales Onion Vs Potato

A pie chart of potato versus onion sales provides essential insights into the sales composition and performance of The business' products. This visualization allows the business to compare product performance, and optimize inventory management which can be seen as onions are in a significant more demand as compared to potatoes. The simplicity and clarity of the pie chart make it easy to understand and support informed decisions in business strategy.



(5.2) Total Sales By Month

The total sales graph for Kanhaiyal D Jaiswal provides essential insights into business performance over a year, aiding in revenue tracking, identifying seasonal trends, and evaluating sales strategies. It helps in budgeting and forecasting by offering historical sales data, which is crucial for setting realistic targets and financial planning. A bar graph was chosen to represent monthly total sales due to its effectiveness in facilitating clear comparisons, highlighting trends, and being easy to interpret. The bar graph "Total Sales by Month" shows significant fluctuations in sales throughout the year. The highest sales were recorded in November, reaching nearly 14 lakh. Other notable peaks were observed in May and June, with sales exceeding 10 lakh each. In contrast, July had the lowest sales, just under 4 lakh, followed by a slow recovery in August and September, where sales ranged between 5 lakh and 6 lakh. Sales remained relatively stable during the first quarter, with January, February, and March each generating around 8 lakh in sales. This data indicates seasonal variations in demand, with a notable surge in sales toward the end of the year and a sharp dip in mid-year.



(5.3) Average Price (per 10kg) Of Onion And Potato By Month

Plotting the price trend for potatoes and onions over 12 months provides essential insights into seasonal price fluctuations and market dynamics. This helps The business understand seasonal variations, enabling them to anticipate price changes and make strategic decisions on pricing and stocking

A line graph was chosen to represent the price trend of potatoes and onions for its clarity, readability, and ability to display data over time. It allows for easy comparison of both products' price trends within the same graph, offering detailed insights into monthly changes and overall trends.

The line graph illustrates the price trends for 10 kg of onions and 10 kg of potatoes over the course of a year. Onion prices experienced significant fluctuations, starting at around ₹200 in January and peaking dramatically in October at over ₹500 per 10 kg, followed by a sharp drop in December to around ₹300. Potato prices, on the other hand, were more stable throughout the year, with slight increases and decreases. They began at approximately ₹150 in January, peaked in June at nearly ₹250, and then steadily declined, staying around ₹150 to ₹200 for the rest of the year. The distinct volatility in onion prices compared to the relatively stable potato prices indicates seasonal or supply-related price variations, especially for onions during the latter half of the year.