DAIRY GOODS PRODUCTION AND SALES ANALYSIS

AHILA SASEENDRAN

November 18, 2024

The Dairy Goods Sales Dataset provides a detailed and comprehensive collection of data related to dairy farms, dairy products, sales, and inventory management. This dataset encompasses a wide range of information, including farm location, land area, cow population, farm size, production dates, product details, brand information, quantities, pricing, shelf life, storage conditions, expiration dates, sales information, customer locations, sales channels, stock quantities, stock thresholds, and reorder quantities.This dataset provides an excellent foundation for operational improvements, strategic planning, and customer insights.

Inventory management data, including stock levels, thresholds, and reorder quantities, is crucial for maintaining a balanced supply chain, minimizing waste, and avoiding stockouts. Additionally, sales information provides insights into transaction details, customer locations, and preferred sales channels (e.g., online, in-store, or wholesale), enabling businesses to understand customer preferences and optimize distribution strategies. This dataset serves as a foundation for analyzing production efficiency, identifying best-selling products, forecasting demand, reducing spoilage, and tailoring marketing strategies, making it indispensable for stakeholders across the dairy supply chain. The dataset also includes production metrics such as production dates, shelf life, and storage conditions, essential for tracking product freshness, ensuring quality, and managing logistics. Product-specific details, such as product types (milk, cheese, yogurt, etc.), brand names, pricing, and quantities, allow for analyzing market performance and profitability.

OBJECTIVE

The objective of Dairy Sales Production Analysis using Python and Pandas is to extract actionable insights and make data-driven decisions to optimize production, inventory, and sales strategies in the dairy industry. Specifically, the goals include:

Monitor Sales Performance:

It is a critical aspect of analyzing dairy sales data, as it helps businesses identify patterns and make informed decisions. By examining the dataset, companies can pinpoint top-performing products, revealing which items generate the highest revenue or have the most consistent sales across different periods. Understanding these product trends allows businesses to focus on maintaining or increasing their production and marketing efforts for high-demand items.

Optimize Inventory Management:

It is essential for maintaining a balance between supply and demand in the dairy industry. By analyzing stock levels, businesses can track the availability of products in real-time, ensuring they meet customer demands without overstocking, which can lead to unnecessary storage costs. Examining shelf life and expiration dates allows companies to prioritize the sale or redistribution of products nearing the end of their freshness, thereby reducing waste and minimizing financial losses from unsold goods.

Evaluate Production Efficiency:

Evaluating production efficiency in the dairy industry involves understanding how farm size, production output, and sales are interconnected to optimize operations. By analyzing the relationship between farm size and production output, businesses can determine whether larger farms yield proportionally higher outputs or if smaller farms are more efficient in terms of production per unit area. This analysis helps identify potential areas for improvement in farm management practices or resource allocation.

Improve Revenue and Profitability:

Improving revenue and profitability in the dairy industry involves a thorough analysis of pricing strategies, revenue contributions, and the performance of different products and sales channels. By examining the pricing of various dairy products, businesses can determine whether their pricing strategies are competitive and aligned with market demand. Analyzing revenue contributions from each product helps identify high-margin items, allowing businesses to focus on promoting or expanding these products to increase overall profitability.

Forecast Demand:

It is a crucial strategy for ensuring that dairy businesses can meet customer needs without overproduction or shortages. By analyzing historical sales data and identifying seasonal patterns, companies can predict future demand trends more accurately. For example, if sales typically increase during certain months (such as holidays or summer), businesses can anticipate the need for higher production and stock levels in advance. Conversely, by understanding periods of lower demand, businesses can reduce production to avoid overstocking and minimize waste. Advanced

forecasting methods can also incorporate other factors like market trends, consumer preferences, and external events that could affect demand.

Enhance Customer Insights:

It is a powerful way to tailor products and marketing strategies to meet the specific needs of different customer segments.By segmenting customers based on factors such as location, purchase behavior, or demographics, businesses can gain a deeper understanding of their target market and make more informed decisions. For example, segmenting by location allows companies to identify regional preferences, enabling them to adjust product offerings or distribution strategies to better cater to local tastes or needs.

Assess Promotional Effectiveness:

Evaluating the impact of discounts or promotions on sales involves analyzing sales data before, during, and after promotional periods to understand how these strategies influence customer purchasing behavior, allowing businesses to optimize future marketing campaigns for greater effectiveness and return on investment.

Waste Reduction:

Minimizing spoilage involves aligning production schedules and inventory turnover with actual demand and product shelf life, ensuring that goods are produced and stocked in quantities that meet customer needs while reducing excess inventory that may lead to waste.

Regional Analysis:

Regional analysis involves comparing performance metrics, such as sales, customer engagement, and operational efficiency, across different geographic locations to uncover trends, identify successful markets, and highlight areas for expansion or improvement in operations.

EXPLORATORY DATA ANALYSIS (EDA)

1.DATA COLLECTION AND UNDERSTANDING

Data can be collected from multiple sources, such as files, databases, APIs, or web scraping. Common file formats include CSV, Excel, and JSON. You can use Pandas to load these files into DataFrames.

2.DATA CLEANING

Data cleaning in Python involves preprocessing raw data to make it suitable for analysis. It ensures that the dataset is free from errors, inconsistencies, and irrelevant information. Below are key steps and corresponding Python code snippets for effective data cleaning.

3.FEATURE ENGINEERING

It is the process of creating, transforming, or selecting features (input variables) from raw data to improve the performance of machine learning models. It involves generating meaningful insights, simplifying complex datasets, and preparing data for analysis.It helps capture the essence of the problem, making patterns more evident to the model. Reduces overfitting by removing irrelevant or redundant features.

4.UNIVARIATE ANALYSIS

It involves analyzing a single variable to understand its distribution, central tendency, dispersion, and outliers. It's a fundamental step in exploratory data analysis (EDA) that helps summarize and visualize the characteristics of individual variables.

5.BIVARIATE ANALYSIS

It examines the relationship between two variables, identifying trends, correlations, or differences between them. It can involve combinations of numerical and categorical variables, using statistical methods and visualizations to uncover insights.

6.MULTIVARIATE ANALYSIS

Its examining relationships among three or more variables simultaneously to uncover complex interactions, patterns, or dependencies. It is a natural extension of univariate and bivariate analysis, often used in exploratory data analysis (EDA), predictive modeling, and hypothesis testing.

7.VISUALISATION

It is to process of representing data graphically using Python libraries. It enables you to better understand patterns, trends, and relationships in the data and communicate insights effectively. Python offers a wide range of libraries for creating static, interactive, and publication-quality visualizations. Each library provides tools for generating various types of plots, charts, and graphs.

DATASET OVERVIEW

The Dairy Goods Sales Dataset provides a detailed and comprehensive collection of data related to dairy farms, dairy products, sales, and inventory management. This dataset encompasses a wide range of information, including farm location, land area, cow population, farm size, production dates, product details, brand information, quantities, pricing, shelf life, storage conditions, expiration dates, sales information, customer locations, sales channels, stock quantities, stock thresholds, and reorder quantities

DATA SET SOURCE : Kaggle

ROWS : 4325

COLUMNS : 23

LOCATION - The geographical location of the dairy farm.

TOTAL LAND AREA - The total land area occupied by the dairy farm.

NUMBER OF COWS - The number of cows present in the dairy farm.

FARM SIZE - The size of the dairy farm.

DATE - The date of data recording.

PRODUCT ID - The unique identifier for each dairy product.

PRODUCT NAME - The name of the dairy product.

BRAND - The brand associated with the dairy product

QUANTITY - The quantity of the dairy product available.

PRICE PER UNIT - The price per unit of the dairy product

TOTAL VALUE - The total value of the available quantity of the dairy product.

SHELF LIFE - The shelf life of the dairy product in days.

STORAGE CONDITION - The recommended storage condition for the dairy product.

PRODUCTION DATE - The date of production for the dairy product.

EXPIRATION DATE - The date of expiration for the dairy product.

QUANTITY SOLD - The quantity of the dairy product sold.

PRICE PER UNIT SOLD - The price per unit at which the dairy product was sold.

TOTAL REVENUE - The approximate total revenue generated from the sale of the dairy product.

CUSTOMER LOCATION - The location of the customer who purchased the dairy product.

SALES CHANNEL - The channel through which the dairy product was sold (Retail, Wholesale, Online).

QUANTITY IN STOCK - The quantity of the dairy product remaining in stock.

MINIMUM STOCK THRESHOLD - The minimum stock threshold for the dairy product.

REORDER QUANTITY - The recommended quantity to reorder for the dairy product

The shelf life and storage conditions for each product can guide inventory management decisions. Dairy products with a shorter shelf life may require more frequent monitoring of stock levels and expedited sales to reduce spoilage. The ability to track specific products and brands can offer valuable insights into product performance, customer preferences, and brand loyalty.Using historical data to predict future sales, helping to optimize production schedules and stock levels. Analyzing product shelf life to minimize waste and ensure products are sold before expiration. Using reorder quantity, stock levels, and minimum stock thresholds to ensure that inventory is always available without overstocking, minimizing spoilage. This dataset provides a holistic view of the dairy farm's operations, from production to sales. By analyzing the relationships between the various columns, valuable insights can be derived for optimizing inventory, enhancing product sales, improving customer targeting, and maximizing overall farm profitability. Regular analysis of the data will help in making data-driven decisions to enhance the efficiency and effectiveness of farm operations and business strategy.

Comparing revenue and sales by sales channel to identify the most profitable channels and improve marketing strategies. Analyzing product shelf life to minimize waste and ensure products are sold before expiration. Grouping customers by location or purchase behavior to tailor marketing strategies and improve customer targeting. Analyzing total revenue and product costs to determine which products are most profitable and which may need price adjustments.

UNDERSTANDING DATA

```
df.shape
df.head(5)
df.tail(5)
df.dtypes
df.describe()
df.info()
df.corr(numeric_only=True)
```

Upon analyzing the dataset, it was observed that it contains 4325 rows and 23 columns. To gain an initial understanding of the data, the following steps were taken:
Head and Tail Examination: The head(5) function was used to inspect the first five rows of the dataset, while the tail(5) function was applied to examine the last five rows. This provided insight into the structure, column names, and a sample of the data values.

Data Type Verification**:** The data types of all columns were checked to ensure they align with the expected formats. For categorical and numerical columns, the appropriate data types were confirmed. The date column, initially detected as an object type, was converted to the datetime format to ensure accurate analysis.

Missing Values and Summary**:** The info() method was used to identify any missing values and determine the data type of each column. This step helped verify the integrity of the dataset before proceeding with the data cleaning process.

```
df.isna().sum()
df.duplicated().sum()
df.describe(include='object')
```

```
for i in df:
  if df[i].dtype=='object':
    print(df[i].value_counts())
    print()
```

Duplicate Check: A check for duplicate rows revealed that there are no duplicate entries in the dataset. These initial checks ensured that the dataset is free from missing or duplicate data, setting a solid foundation for the subsequent stages of data cleaning and analysis.

These exploratory steps laid the foundation for the subsequent data cleaning phase by ensuring the data is well-structured and ready for further analysis.

Descriptive Statistics for Categorical Data**:** The describe(include='object') function was used to generate descriptive statistics specifically for categorical columns. This provides insights into the distribution of values, including the count, unique values, top (most frequent) value, and frequency.

Value Counts for Categorical Columns**:** A loop was implemented to iterate through all columns. For columns with a data type of 'object' (i.e., categorical variables), the value_counts() method.

DATA CLEANING

df['Date']=pd.to_datetime(df['Date'])
df['Production Date']=pd.to_datetime(df['Production Date'])
df['Expiration Date']=pd.to_datetime(df['Expiration Date'])
df.rename(columns={'Total Land Area (acres)': 'Total Land Area'}, inplace=True)

Purpose**:** These lines convert the columns 'Date', 'Production Date', and 'Expiration Date' from object type (likely string) to datetime type.

Benefit**:** Ensures that the date-related columns are correctly recognized as date objects, which allows for date-specific operations such as filtering by date, calculating date differences, and performing time series analysis.

df.rename(columns={'Quantity (liters/kg)': 'Quantity'}, inplace=True)
df.rename(columns={'Shelf Life (days)': 'Shelf Life'}, inplace=True)
df.rename(columns={'Quantity Sold (liters/kg)': 'Quantity Sold'}, inplace=True)
df.rename(columns={'Price per Unit (sold)': 'Price per Unit Sold'}, inplace=True)
df.rename(columns={'Approx. Total Revenue(INR)': 'Total Revenue'}, inplace=True)
df.rename(columns={'Minimum Stock Threshold (liters/kg)': 'Minimum Stock Threshold'}, inplace=True)
df.rename(columns={'Quantity in Stock (liters/kg)': 'Quantity in Stock'}, inplace=True)
df.rename(columns={'Reorder Quantity (liters/kg)': 'Reorder Quantity'}, inplace=True)

Purpose: These lines rename several columns in the dataset to make them more concise, readable, and standardized. The previous names were likely longer or included unnecessary units (e.g., 'liters/kg', 'INR') that are not required in the column name itself.

Benefit: The new column names are clearer and more concise, making the dataset easier to work with and understand. For example:

'Total Land Area (acres)' is renamed to 'Total Land Area'

'Quantity (liters/kg)' is renamed to 'Quantity'

'Approx. Total Revenue(INR)' is renamed to 'Total Revenue'

This standardization ensures consistency across the dataset and helps avoid confusion when referencing the columns in future analyses. The purpose of renaming several columns in the dataset is to make them more concise, readable, and standardized by removing unnecessary units (e.g., 'liters/kg', 'INR') and simplifying lengthy descriptions. This enhances clarity and ensures consistency across the dataset, making it easier to work with and understand. For instance, 'Total Land Area (acres)' becomes 'Total Land Area', and 'Quantity (liters/kg)' is shortened to 'Quantity'.

FEATURE ENGINEERING

```
def Sales_Metric(Quantity_Sold):
  if Quantity_Sold<100:
    return "Low"
  elif Quantity_Sold<500 and Quantity_Sold<1000:
    return "Average"
  else:
    return "High"
df['Sales_Metric'] = df['Quantity Sold'].apply(Sales_Metric)
df
```

It becomes a derived feature that can help improve the performance of your model by providing meaningful categorical information based on sales data. This transformation helps capture important sales trends in terms of low, average, or high performance. The Sales_Metric column is created through a custom function that categorizes the quantity of items sold into three distinct levels: Low, Average, and High. This feature provides valuable insight into the sales performance.

Low Sales Metric: When the quantity sold is less than 100, this indicates low sales activity.

Average Sales Metric: When the quantity sold is between 500 and 1000, it suggests average sales performance.

High Sales Metric: When the quantity sold exceeds 1000, it reflects high sales performance.

This feature engineering step allows to transform raw sales data into a more usable and meaningful feature for machine learning models, improving model interpretability and performance.

OUTLIER DETECTORS

```
q1=df['Price per Unit Sold'].quantile(0.25)
q3=df['Price per Unit Sold'].quantile(0.75)
iqr=q3-q1
min_range=q1-1.5*iqr
max_range=q3+1.5*iqr
df[(df['Price per Unit Sold'] <= max_range) & (df['Price per Unit Sold'] >= min_range)]
df
```

This code filters out outliers in the 'Price per Unit Sold' column of the DataFrame by calculating the Interquartile Range (IQR), determining the lower and upper bounds for valid values, and then selecting only the rows where the price is within those bounds. By removing the outliers, its ensured that the data is cleaner, which can lead to more accurate analysis and model performance. Any rows where the 'Price per Unit Sold' falls outside of this range (i.e., outliers) are excluded.

The first quartile (Q1) represents the 25th percentile of the data, which is the value below which 25% of the data points fall. The third quartile (Q3) represents the 75th percentile, where 75% of the data points fall below this value.

UNIVARIATE ANALYSIS

Univariate analysis of dairy goods sales focuses on examining the distribution and characteristics of sales data for dairy products. This analysis helps in understanding the sales patterns, identifying trends, and detecting potential anomalies in the data.

```
df['Location'].value_counts()

df['Product Name'].value_counts()

df['Brand'].value_counts()

for i in df:
  if df[i].dtype=='object':
    print(df[i].value_counts())
    print()
```

The analysis reveals that the dairy farm located in Delhi has the highest value counts, while Tamil Nadu records the lowest value counts.

```
Location
Delhi            525
Chandigarh       519
Uttar Pradesh    276
Gujarat          267
Karnataka        261
Madhya Pradesh   259
Rajasthan        256
Maharashtra      255
Haryana          253
Kerala           249
Telangana        248
Jharkhand        248
Bihar            245
West Bengal      241
Tamil Nadu       223
Name: count, dtype: int64
```

The farm sizes are categorized into three groups based on their dimensions : small, medium, and large.

```
Farm Size
Large     1462
Medium    1439
Small     1424
Name: count, dtype: int64
```

The analysis indicates that the product “Curd" has the highest value count, while "Cheese" has the lowest.

```
Product Name
Curd          479
Lassi         447
Paneer        441
Yogurt        437
Buttermilk    435
Butter        431
Milk          429
Ice Cream     423
Ghee          402
Cheese        401
Name: count, dtype: int64
```

The analysis indicates that the brand "Amul" has the highest value count, while "Britannia Industries" has the lowest.

```
Brand
Amul                   1053
Mother Dairy           1010
Raj                     685
Sudha                   648
Dodla Dairy             222
Palle2patnam            211
Dynamix Dairies         106
Warana                  104
Parag Milk Foods        102
Passion Cheese           96
Britannia Industries     88
Name: count, dtype: int64
```

The analysis reveals that the most commonly used storage condition for preserving dairy products is refrigeration, while the least utilized is tetra pack.

```
Storage Condition
Refrigerated       2459
Frozen             1035
Ambient             402
Polythene Packet    225
Tetra Pack          204
Name: count, dtype: int64
```

The analysis indicates that the highest customer demand for dairy products is in Delhi, while the lowest demand is in Haryana.

```
Customer Location
Delhi            499
Chandigarh       489
Bihar            284
Maharashtra      271
Kerala           267
Tamil Nadu       267
Uttar Pradesh    267
Karnataka        264
West Bengal      264
Telangana        251
Madhya Pradesh   248
Gujarat          248
Jharkhand        243
Rajasthan        234
Haryana          229
Name: count, dtype: int64
```

The analysis shows that the highest sales count is for retail sales, while the lowest sales count is for online sales.

```
Sales Channel
Retail      1478
Wholesale   1476
Online      1371
Name: count, dtype: int64
```

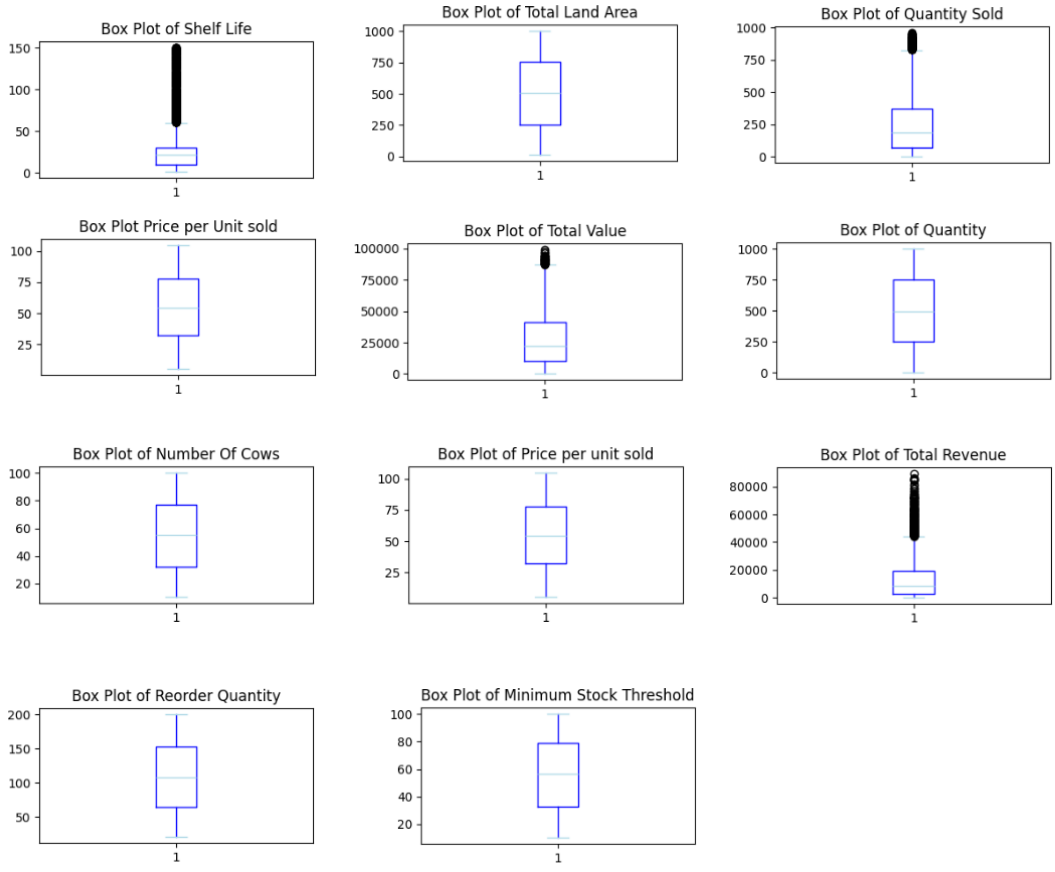
By analysing through the sales metric as average.medium, high through analysis there is only average sales.

```
Sales_Metric
Average    2263
Low        1414
High        648
Name: count, dtype: int64
```

BOX PLOT

It is a standardized way of displaying the distribution of a dataset. It provides a five-number summary of the data, making it easy to identify key characteristics like spread, central tendency, and potential outliers.

```
plt.figure(figsize=(5, 3))
import matplotlib.pyplot as plt
plt.boxplot(df['Quantity'],
        boxprops=dict(color='blue'),
        medianprops=dict(color='lightblue'),
        whiskerprops=dict(color='blue'),
        capprops=dict(color='lightblue'))
plt.title('Box Plot of Quantity')
plt.show()
```



```
for col in df:
  if df[col].dtype=='int' or df[col].dtype=='float':
    print(col)
    print('_____________________')
    print(f'Mean :{df[col].mean ()}')
    print(f'Median :{df[col].median ()}')
    print(f'Minimum :{df[col].min ()}')
    print(f'Maximum :{df[col].max ()}')
    print(f'Standard Deviation :{df[col].std ()}')
    print(f'Variance :{df[col].var ()}')
    print()
    print()
    print()
```

```
Total Land Area
_____________
Mean :503.48307283236994
Median :509.17
Minimum :10.17
Maximum :999.53
Standard Deviation :285.9350614091884
Variance :81758.85934307633

Product ID
_____________
Mean :5.509595375722544
Median :6.0
Minimum :1
Maximum :10
Standard Deviation :2.842978748863798
Variance :8.082528166491167

Quantity
_____________
Mean :500.6526566473989
Median :497.55
Minimum :1.17
Maximum :999.93
Standard Deviation :288.97591543272085
Variance :83507.07970017905

Price per Unit
_____________
Mean :54.78593757225434
Median :54.4
Minimum :10.03
Maximum :99.99
Standard Deviation :26.00281475158703
Variance :676.1463750053521

Number of Cows
_____________
Mean :54.963699421965316
Median :55.0
Minimum :10
Maximum :100
Standard Deviation :26.111486678134206
Variance :681.80973654238

Shelf Life
_____________
Mean :29.12763005780347
Median :22.0
Minimum :1
Maximum :150
Standard Deviation :30.272114446989033
Variance :916.4009130916021
```

```
Quantity Sold
_____________
Mean :248.0950289017341
Median :189.0
Minimum :1
Maximum :960
Standard Deviation :217.02418151325688
Variance :47099.49536149907
```

```
Price per Unit Sold
_____________
Mean :54.779139884393054
Median :54.14
Minimum :5.21
Maximum :104.51
Standard Deviation :26.192790484369684
Variance :686.0622733580872
```

```
Quantity in Stock
_____________
Mean :252.06867052023122
Median :191.0
Minimum :0
Maximum :976
Standard Deviation :223.62086974805584
Variance :50006.293386876954
```

```
Total Revenue
_____________
Mean :13580.265401156072
Median :8394.54
Minimum :12.54
Maximum :89108.9
Standard Deviation :14617.009122450516
Variance :213656955.6858016
```

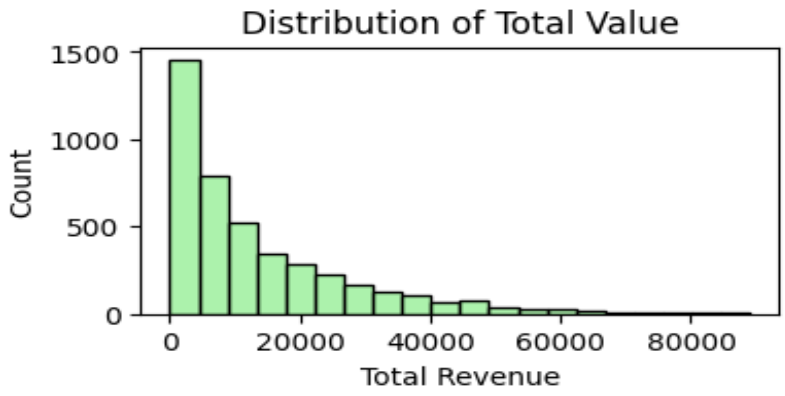
```
Minimum Stock Threshold
_____________
Mean :55.82614335260116
Median :56.46
Minimum :10.02
Maximum :99.99
Standard Deviation :26.3014497665468
Variance :691.7662598221847
```

```
Reorder Quantity
_____________
Mean :109.10781965317919
Median :108.34
Minimum :20.02
Maximum :199.95
Standard Deviation :51.501035148832884
Variance :2652.3566214013204
```

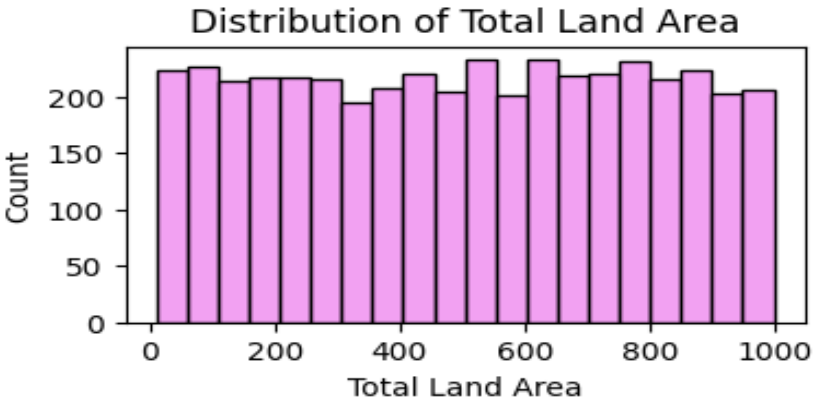
HISTPLOT

It is a graphical representation of distribution of numerical data.It divides data into bins or intervals and displays the frequency of data points in each bin.

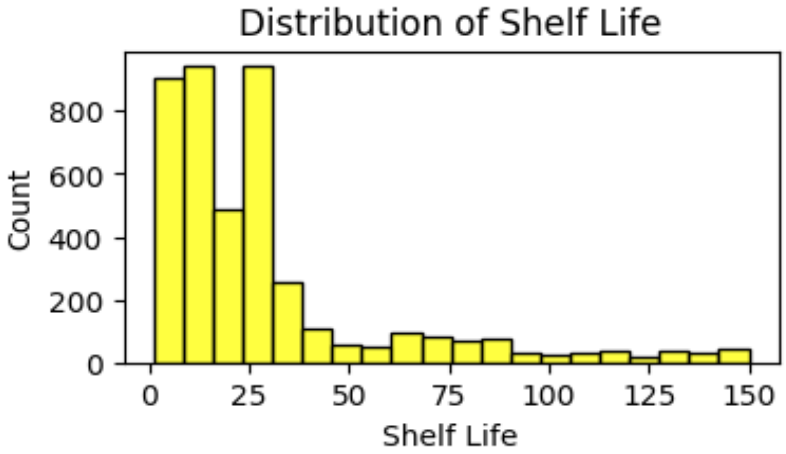
```
plt.figure(figsize=(4, 2))
import seaborn as sns
sns.histplot(data=df, x='Total Revenue', bins=20, color='lightgreen')
plt.title('Distribution of Total Value')
plt.show()
```



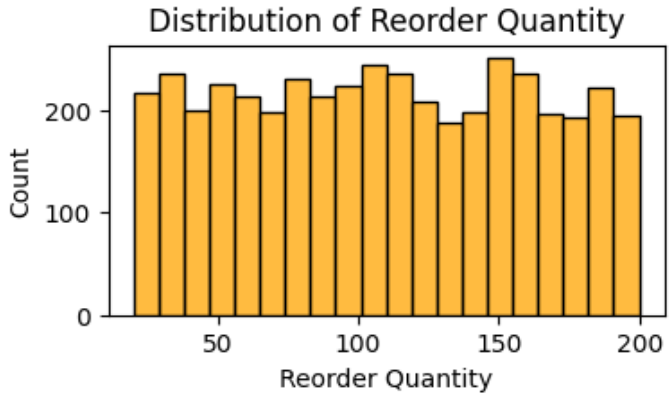
```
plt.figure(figsize=(4, 2))
sns.histplot(data=df, x='Total Land Area', bins=20,color='violet')
plt.title('Distribution of Total Land Area')
plt.show()
```




```
plt.figure(figsize=(4, 2))
sns.histplot(data=df, x='Shelf Life', bins=20,color='yellow')
plt.title('Distribution of Shelf Life')
plt.show()
```




```
plt.figure(figsize=(4, 2))
sns.histplot(data=df, x='Reorder Quantity', bins=20,color='orange')
plt.title('Distribution of Reorder Quantity')
plt.show()
```



Based on the analysis conducted using a histogram plot, it was observed that the majority of revenue values fall within the range of 0 to 20,000, with approximately 1,500 occurrences recorded in this range. This indicates a significant concentration of revenue within this interval, suggesting it represents the most frequent revenue category in the dataset.

The analysis of land area distribution reveals that the range of 400 to 800 exhibits the highest frequency, with a count exceeding 200. This range represents the most prevalent land area category in the dataset, indicating that a significant portion of the observations falls within this interval.

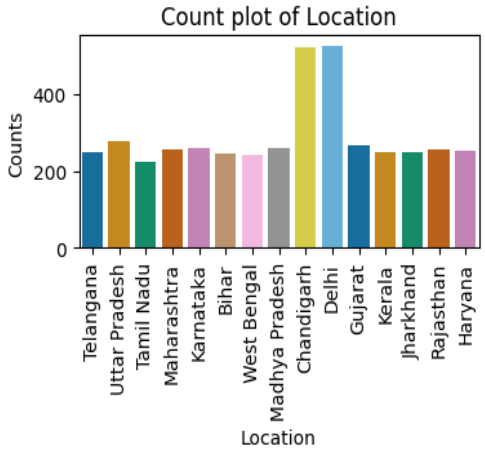
The analysis of product shelf life indicates that the 0-25 days range has the highest count of products. This suggests that a significant proportion of items in the dataset fall within this short shelf-life category

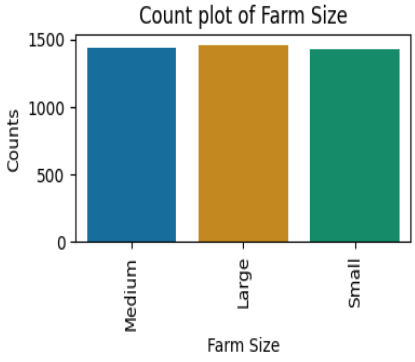
The analysis reveals that 150 products have been reordered with a frequency exceeding 200 counts. This indicates a subset of high-demand items that consistently require replenishment, highlighting their critical role in the product portfolio.

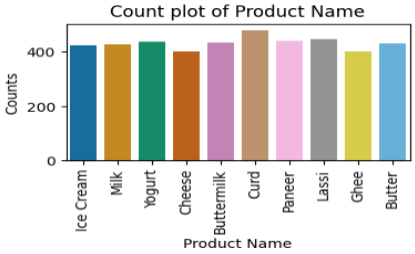
COUNTPLOT

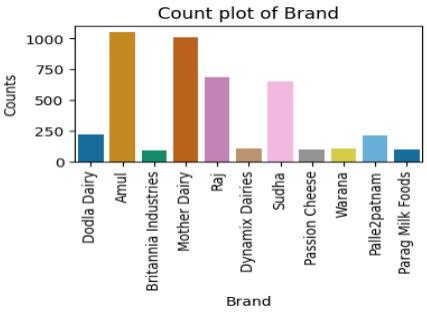
It is a type of categorical plot that displays the count of occurrences for each unique value in a categorical dataset.

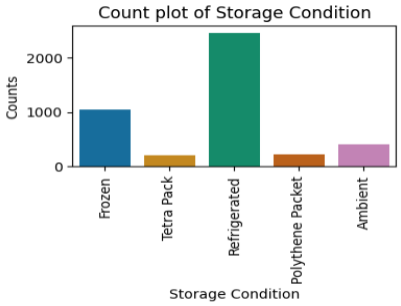
```
for col in df:
   if df[col].dtype == 'object':
      plt.figure(figsize=(4, 2))
      sns.countplot(x=df[col], hue=df[col], palette='colorblind',
dodge=False, legend=False)
      plt.xlabel(col)
      plt.ylabel('Counts')
      plt.xticks(rotation=90)
      plt.title(f'Count plot of {col}') plt.show()
```

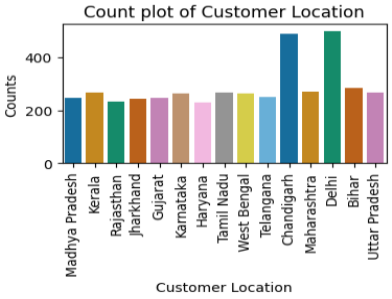


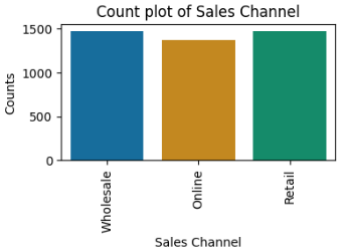












The count plot analysis reveals that the majority of farms are concentrated in Delhi and Chhattisgarh, indicating these regions have the highest representation in the dataset. Conversely, Tamil Nadu exhibits the lowest number of farms, highlighting a relatively limited presence in this region

The analysis indicates that large farm sizes exhibit a higher count, suggesting that a significant proportion of the farms in the dataset have larger land areas. This could imply that large-scale farming operations are more prevalent in the region or dataset being analyzed.

The analysis indicates that curd is the most abundant product in the dataset, with the highest quantity produced, whereas cheese represents the least produced product. This distribution suggests that curd is a more commonly produced or consumed item, possibly due to its widespread popularity, lower production costs, or easier processing compared to cheese

The analysis indicates that Amul is the most popular brand, with the highest product count in the dataset, suggesting that Amul has a dominant market presence and is widely preferred by consumers. This could be due to its strong brand recognition, extensive distribution network, and broad range of dairy products that cater to diverse customer needs. Britannia Industries shows the least product count, indicating a relatively lower market share or product availability compared to Amul.

The analysis reveals that the most preferred storage condition for dairy products is the refrigerator, suggesting that consumers and businesses alike prioritize maintaining the freshness and quality of these products by storing them in cold conditions. This preference is likely due to the perishable nature of dairy items, such as milk, cheese, and yogurt, which require consistent refrigeration to prevent spoilage and maintain safety.

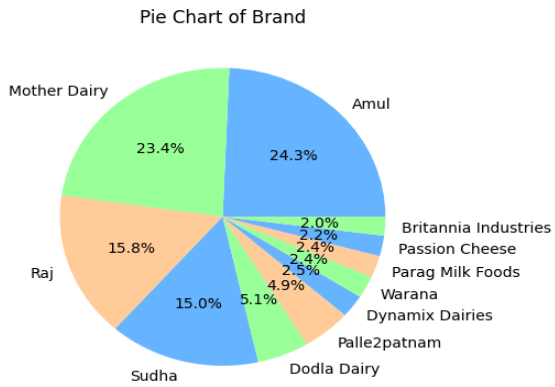
The analysis indicates that Delhi is both the region with the highest customer interest in dairy products and the location with the largest number of farms dedicated to dairy production. This dual trend suggests a strong demand for dairy products in Delhi, likely driven by its large population and urbanization, which supports a consistent need for dairy-based goods.Mostly the retail sales takes place more than whole sale and online

The analysis reveals that the sales figures for dairy products are relatively average, with most sales counts falling within the range of 5,000 units.

PIE PLOT

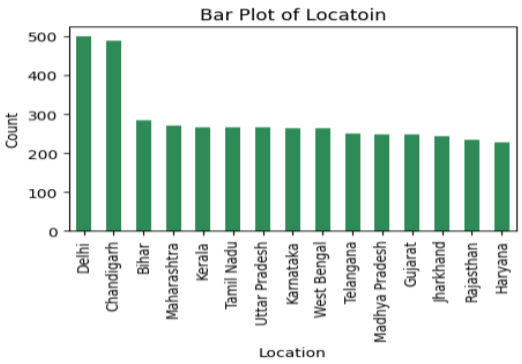
A pie plot in Pandas is a way to display the distribution of categorical data in the form of a pie chart. It is useful for showing proportions or percentages for a particular column in your dataset.

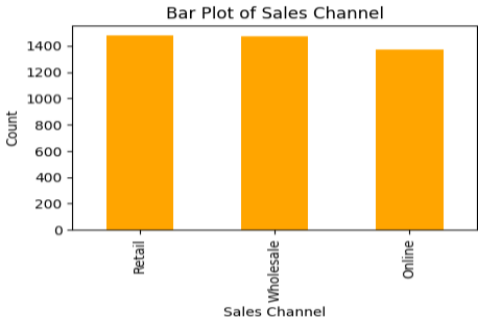
```
df['Brand'].value_counts().plot(kind='pie', autopct='%1.1f%%',
colors=['#66b3ff','#99ff99','#ffcc99'])
plt.ylabel('')
plt.title('Pie Chart of Column Name')
plt.show()

df['Product Name'].value_counts().plot(kind='pie', autopct='%1.1f%%',
colors = ['#FFA07A', '#20B2AA', '#9370DB', '#FF6347', '#4682B4'])
plt.ylabel('')  # To remove the y-label
plt.title('Pie Chart of Column Name')
```

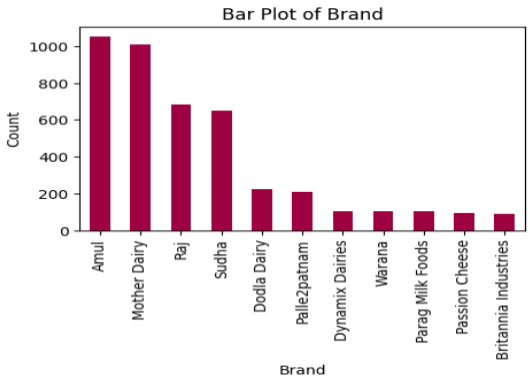


BAR PLOT

A bar plot is a type of data visualization used to represent categorical data with rectangular bars. The length or height of each bar is proportional to the value or frequency of the category it represents. Bar plots are often used to compare different categories or groups within a dataset.







The univariate analysis provides a strong foundation for further exploration and decision-making in the dairy industry. Key insights, such as product preferences, regional production trends, and storage conditions, offer a clear picture of the current state of the market. Moving forward, these insights can be used to optimize production strategies, marketing efforts, and supply chain logistics, ensuring better alignment with consumer demand and regional variations.

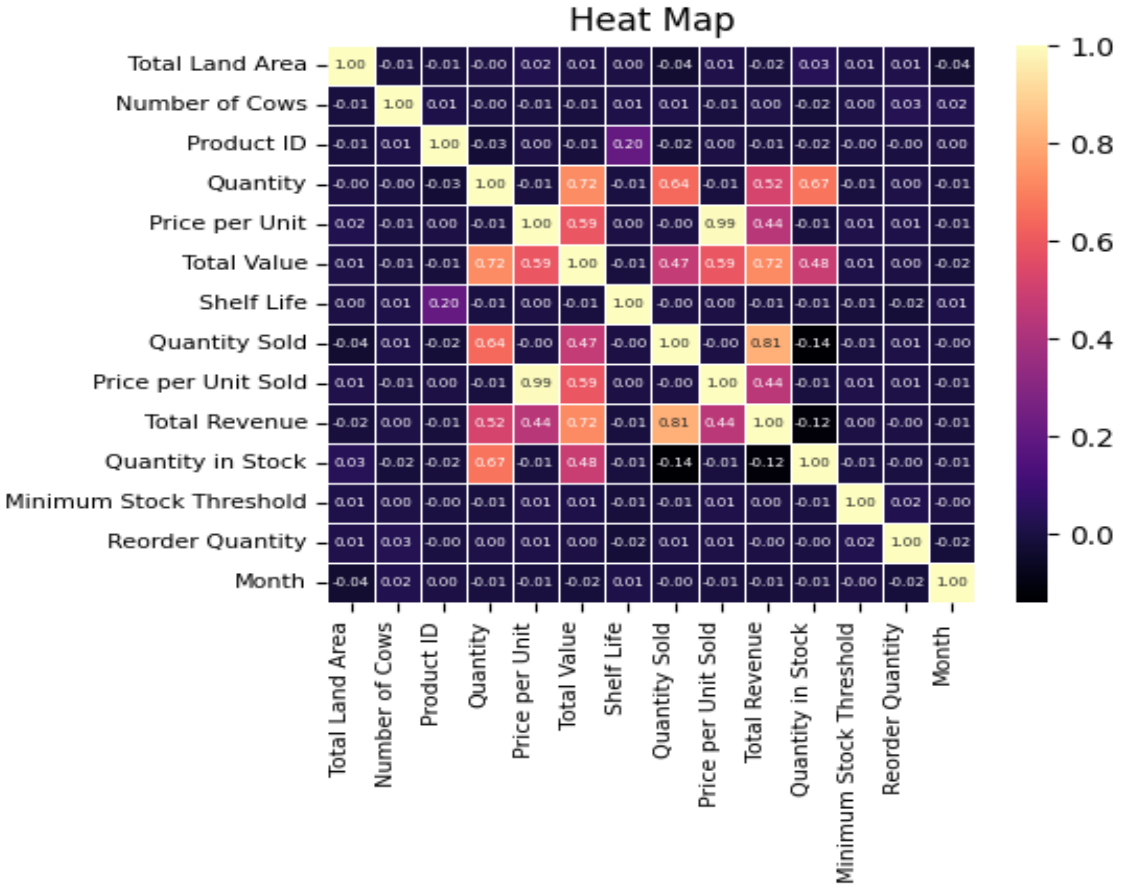
BIVARIATE ANALYSIS

Bivariate analysis involves the analysis of two variables to understand the relationship between them. It helps in identifying correlations, dependencies, and trends between two variables in a dataset. In Python, bivariate analysis can be performed using various techniques like scatter plots, correlation matrices, box plots, and line plots

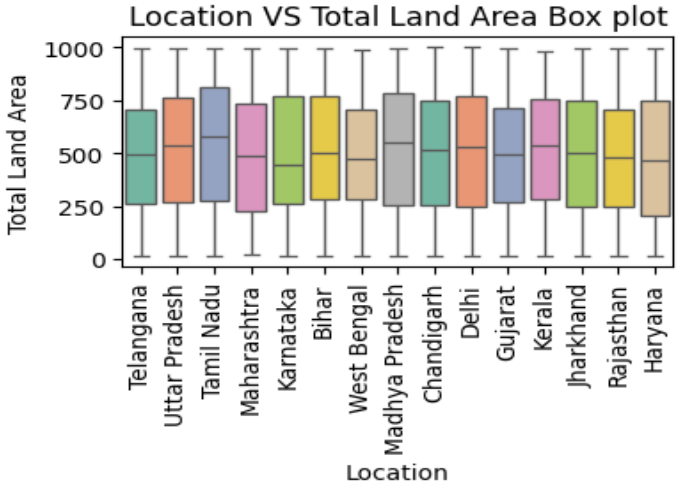
HEAT MAP

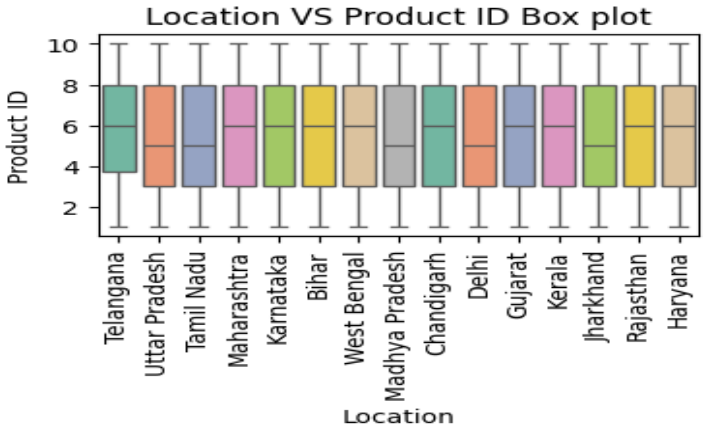
Heat Map is to show the relationship between two variables, The gradient represents the magnitude or frequency of the relationship between the two variables. Darker color shows negative correlation and lighter color shows positive correlation

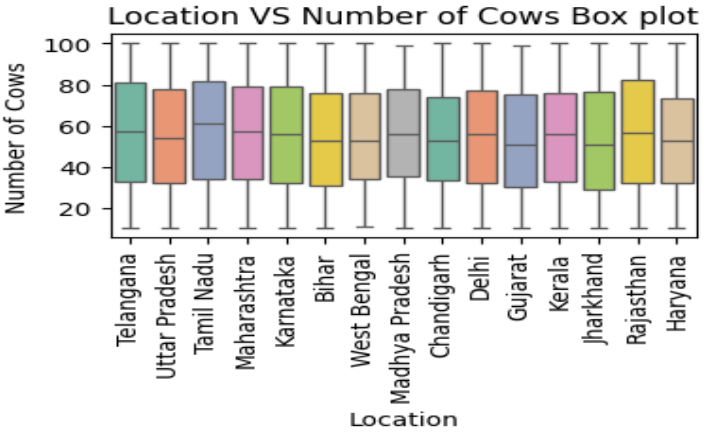
```
var1 = df.corr(numeric_only=True)
plt.figure(figsize=(5, 4))  # Adjust the figure size as needed
sns.heatmap(var1, annot=True, fmt=".2f", annot_kws={"size": 5},
cmap='magma', linewidths=0.5)
plt.title('Heat Map')
plt.xticks(rotation=90, ha='right', fontsize=8)
plt.yticks(rotation=0, fontsize=8)
plt.show()
```



BOX PLOT OF CATEGORICAL AND NUMERICAL COLUMN







Location VS Quantity Box plot
Quantity
1000
750
500
250
0
Telangana
Uttar Pradesh
Tamil Nadu
Maharashtra
Karnataka
Bihar
West Bengal
Madhya Pradesh
Chandigarh
Delhi
Gujarat
Kerala
Jharkhand
Rajasthan
Haryana
Location

Location VS Price per Unit Box plot
Price per Unit
100
80
60
40
20
Telangana
Uttar Pradesh
Tamil Nadu
Maharashtra
Karnataka
Bihar
West Bengal
Madhya Pradesh
Chandigarh
Delhi
Gujarat
Kerala
Jharkhand
Rajasthan
Haryana

Location VS Total Value Box plot
Total Value
100000
75000
50000
25000
0
Telangana
Uttar Pradesh
Tamil Nadu
Maharashtra
Karnataka
Bihar
West Bengal
Madhya Pradesh
Chandigarh
Delhi
Gujarat
Kerala
Jharkhand
Rajasthan
Haryana
Location

Location VS Shelf Life Box plot
Shelf Life
150
100
50
0
Telangana
Uttar Pradesh
Tamil Nadu
Maharashtra
Karnataka
Bihar
West Bengal
Madhya Pradesh
Chandigarh
Delhi
Gujarat
Kerala
Jharkhand
Rajasthan
Haryana
Location

```
df.sort_values(by='Quantity Sold',ascending=False).head(100)
df.sort_values(by='Quantity Sold',ascending=True).head(41)
df.groupby('Location')['Quantity'].mean().sort_values(ascending=False)
df.groupby('Location')['Total Land
Area'].mean().sort_values(ascending=False)
df.groupby('Location')['Number of
Cows'].mean().sort_values(ascending=False)
df.groupby('Farm Size')['Total Land
Area'].mean().sort_values(ascending=False)
df.groupby('Product
Name')['Quantity'].mean().sort_values(ascending=False)
df.groupby('Product Name')['Reorder
Quantity'].mean().sort_values(ascending=False)
result
df.groupby('Brand')['Reorder
Quantity'].mean().sort_values(ascending=False)
df.groupby('Product Name')['Shelf
Life'].mean().sort_values(ascending=False)
df.groupby('Month')['Quantity
Sold'].mean().sort_values(ascending=False)
df.groupby('Quantity in Stock')['Reorder
Quantity'].mean().sort_values(ascending=False)
df.groupby('Product Name')['Price per
Unit'].mean().sort_values(ascending=False)
df.groupby('Customer Location')['Reorder
Quantity'].mean().sort_values(ascending=False)
```

The analysis reveals that the highest quantity of products is sold from the farm located in Tamil Nadu, with the customer location primarily being Telangana. This indicates a strong supply-demand connection between these regions, highlighting Tamil Nadu's significant contribution to meeting Telangana's dairy product needs. This insight could be leveraged to optimize logistics and strengthen the supply chain between the two states, ensuring sustained customer satisfaction and operational efficiency.

The analysis indicates that Tamil Nadu has the highest farm area, emphasizing its prominence in agricultural and dairy production. Conversely, Haryana is identified as having the least farm area, suggesting limited land allocation for such activities in the region. These findings highlight regional variations in land use and production capacity, which could influence strategic decisions related to resource allocation and operational planning.

The highest number of cows is located on farms in Tamil Nadu, signifying its dominant role in dairy production. On the other hand, Gujarat has the least number of cows on its farms, suggesting a smaller scale of dairy farming activities in the region.

The analysis reveals that curd is the product with the highest production quantity, indicating its strong demand and widespread preference among consumers. In contrast, paneer is produced in significantly lower quantities.

Milk has the highest reorder quantity, highlighting its consistent demand and essential role in consumer purchases. Conversely, ice cream has the lowest reorder quantity.

Parag Milk Foods has the highest reorder quantity, underscoring its popularity and consistent demand in the market. On the other hand, Passion Cheese has the lowest reorder quantity, indicating a more limited market presence or niche customer base.

Ghee has the longest shelf life, making it suitable for extended storage and distribution. In contrast, curd has the shortest shelf life, requiring prompt consumption and careful handling.

The highest quantity of products is sold in August, reflecting peak demand during this month, possibly due to seasonal trends or festivals. Conversely, April records the lowest quantity sold, which may be attributed to reduced market activity or lower seasonal demand.

The quantity in stock consistently exceeds the reorder quantity, indicating a well-maintained inventory buffer to meet demand fluctuations.

Reorder quantity is higher at location Gujarat and less at uttarpradesh.

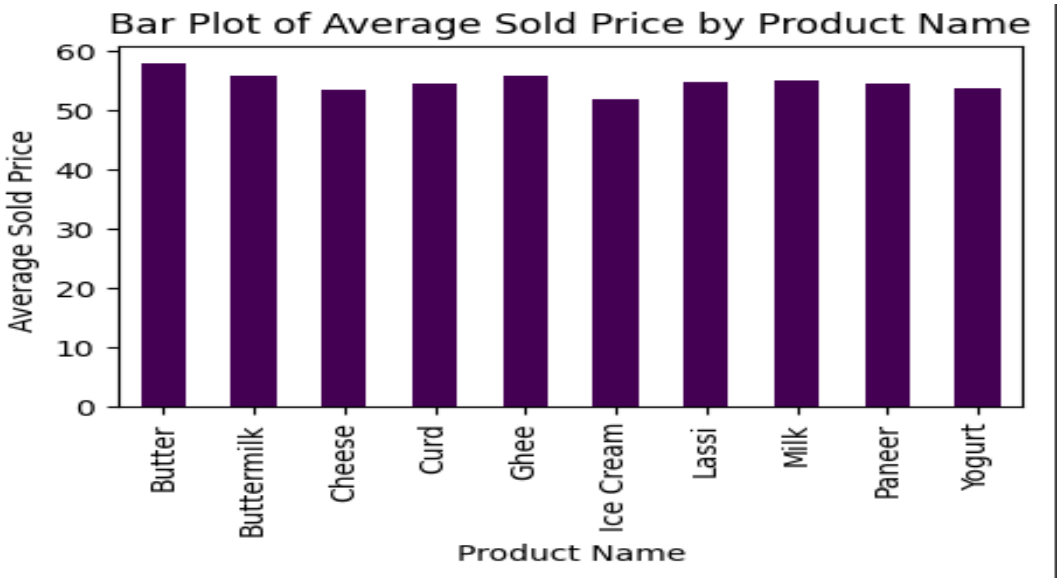
df[['Brand','Minimum Stock Threshold']].groupby('Brand').agg(['mean','median','min','max','std','count','var']).sort_values(by=('Minimum Stock Threshold','mean'),ascending=False)

	Minimum Stock Threshold						
	mean	median	min	max	std	count	var
Brand							
Palle2patnam	57.846256	59.250	10.03	99.92	24.802593	211	615.168623
Raj	57.376496	57.830	10.17	99.99	26.342452	685	693.924774
Amul	56.825499	58.230	10.10	99.97	26.162129	1053	684.457008
Warana	56.194904	56.055	11.16	99.62	26.746725	104	715.387316
Britannia Industries	56.002159	56.600	10.21	98.67	27.643928	88	764.186748
Mother Dairy	55.508594	56.490	10.19	99.88	26.104491	1010	681.444451
Dodla Dairy	55.426802	55.470	10.59	99.85	27.768695	222	771.100444
Passion Cheese	53.916250	54.555	11.47	99.74	27.304747	96	745.549184
Sudha	53.531435	52.460	10.02	99.96	26.492058	648	701.829138
Dynamix Dairies	53.485943	55.635	10.09	96.35	26.062786	106	679.268834
Parag Milk Foods	53.212059	52.320	10.07	99.40	25.206432	102	635.364226

df[['Product Name','Price per Unit']].groupby('Product Name').agg(['mean','median','std','min','max','count']).sort_values(by=('Price per Unit', 'mean'),ascending=False)

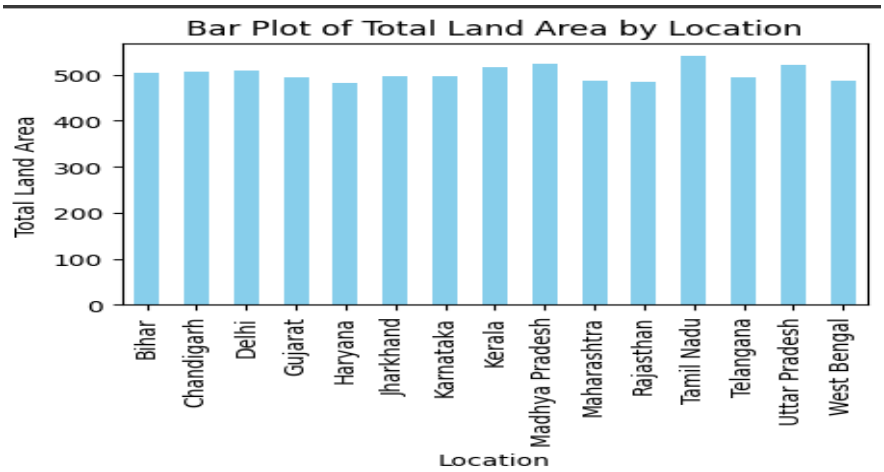
	Price per Unit					
	mean	median	std	min	max	count
Product Name						
Butter	57.796984	59.92	26.295910	10.61	99.96	431
Ghee	55.907040	57.41	25.898104	10.19	99.65	402
Buttermilk	55.863770	57.51	25.947067	10.40	99.99	435
Milk	54.933403	53.89	26.522287	10.09	99.70	429
Curd	54.861106	55.60	25.562998	10.33	99.78	479
Paneer	54.717551	53.50	26.400326	10.11	99.49	441
Lassi	54.672036	53.39	26.436229	10.03	99.30	447
Yogurt	53.876613	51.69	25.794228	10.16	99.51	437
Cheese	53.505387	54.19	25.575131	10.05	99.96	401
Ice Cream	51.654444	50.17	25.367146	10.16	99.73	423

df.groupby('Product Name')['Price per Unit Sold'].mean().plot(kind='bar', colormap='viridis', figsize=(5,3))
plt.xlabel('Product Name')
plt.ylabel('Average Sold Price ')
plt.title('Bar Plot of Average Sold Price by Product Name')
plt.show()



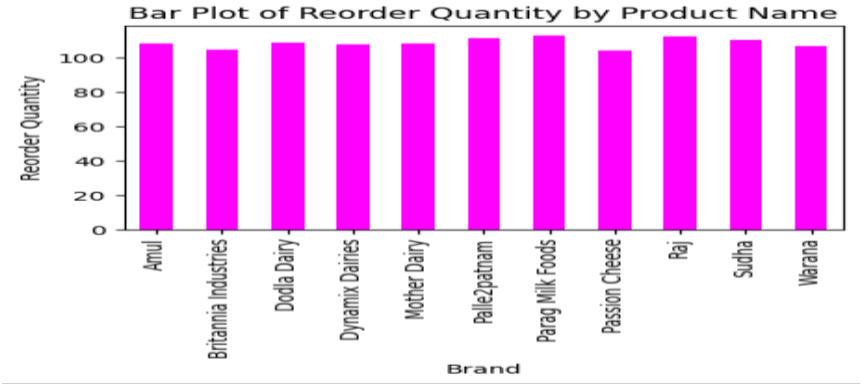
The average sold price of butter is higher compared to that of ice cream, as shown in the bar plot. This suggests a price disparity between the two products, which could be further explored for strategic pricing decisions.

```
df.groupby('Location')['Total Land Area'].mean().plot(kind='bar',
color='green', figsize=(5,3))
plt.xlabel('Location')
plt.ylabel('Total Land Area')
plt.title('Bar Plot of Total Land Area by Location')
plt.show()
```



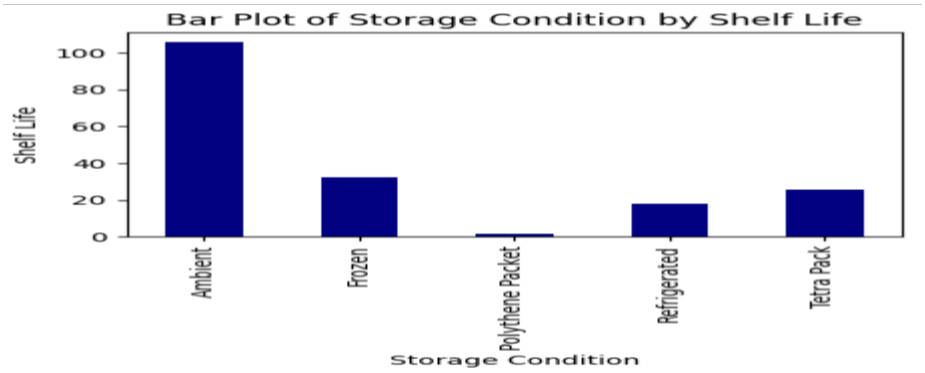
Tamil Nadu has the highest land area among the regions analyzed. Additionally, from previous analyses, it was found that Tamil Nadu also hosts the largest number of cows, suggesting a potential correlation between land area and livestock population.

```
df.groupby('Brand')['Reorder Quantity'].mean().plot(kind='bar',
colormap='spring', figsize=(5, 3))
plt.xlabel('Brand')
plt.ylabel('Reorder Quantity')
plt.title('Bar Plot of Reorder Quantity by Product Name')
plt.show()
```



Parag Milk products have the highest reorder quantity, indicating strong demand and inventory turnover. In contrast, Passion Cheese has the lowest reorder quantity, suggesting lower demand or differing inventory strategies.

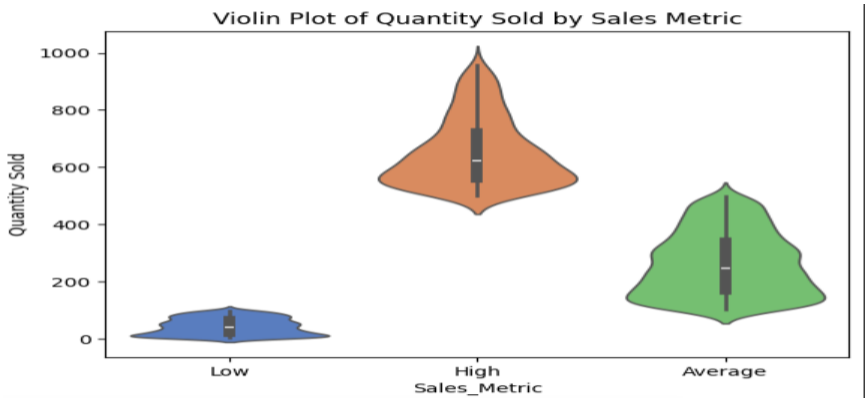
```
df.groupby('Storage Condition')['Shelf Life'].mean().plot(kind='bar',
colormap='jet', figsize=(5, 3))
plt.xlabel('Storage Condition')
plt.ylabel('Shelf Life')
plt.title('Bar Plot of Storage Condition by Shelf Life')
plt.show()
```



Ambient storage conditions are effective in extending shelf life. However, refrigerated storage is the most preferred option as it ensures dairy products remain fresh for a longer duration.

VIOLIN PLOT

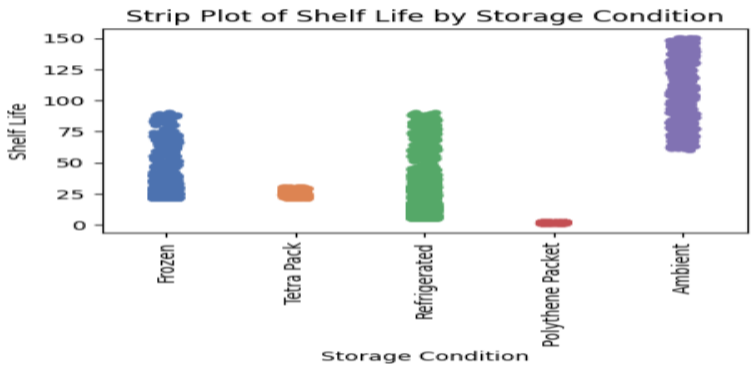
It is a data visualization tool used to represent the distribution of a dataset. It combines aspects of a box plot and a kernel density plot, providing detailed information about the distribution's shape, spread, and central tendency.



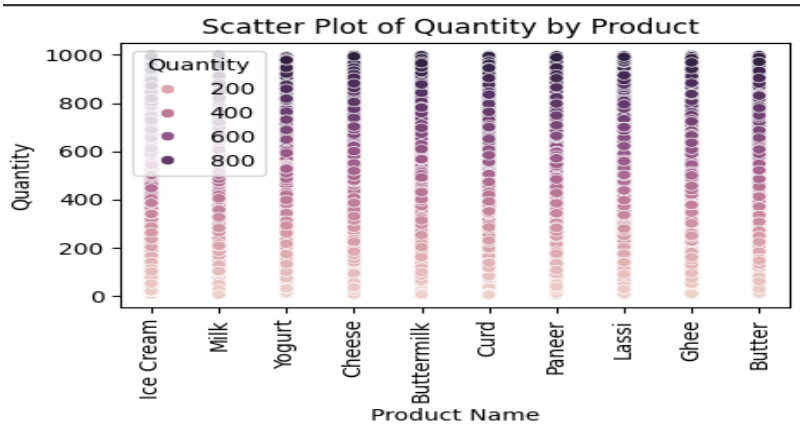
Sales quantities exceeding 500 are considered a high-performance metric, as visualized using a violin plot. This representation highlights the distribution and density of sales data within this range.

STRIP PLOT

It is a simple scatter plot used to display the distribution of a dataset along a single categorical axis. It plots individual data points, making it useful for visualizing data distributions and potential outliers.

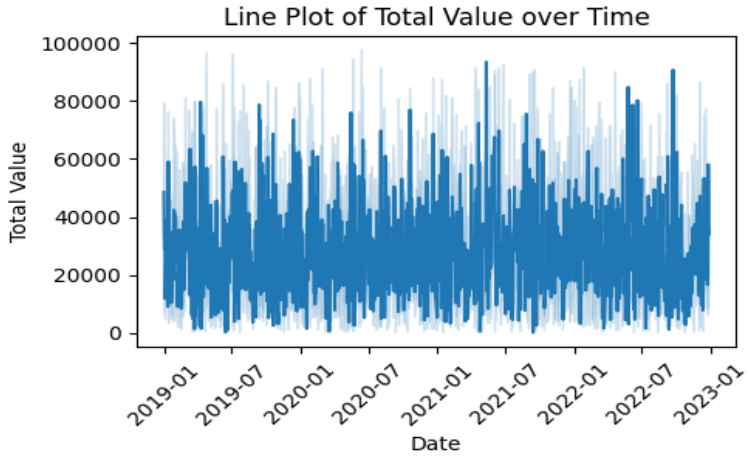



```
sns.scatterplot(data=df, x='Product Name', y='Quantity', hue='Quantity')
plt.xticks(rotation=90)
plt.title("Scatter Plot of Quantity by Product ")
plt.show()
```

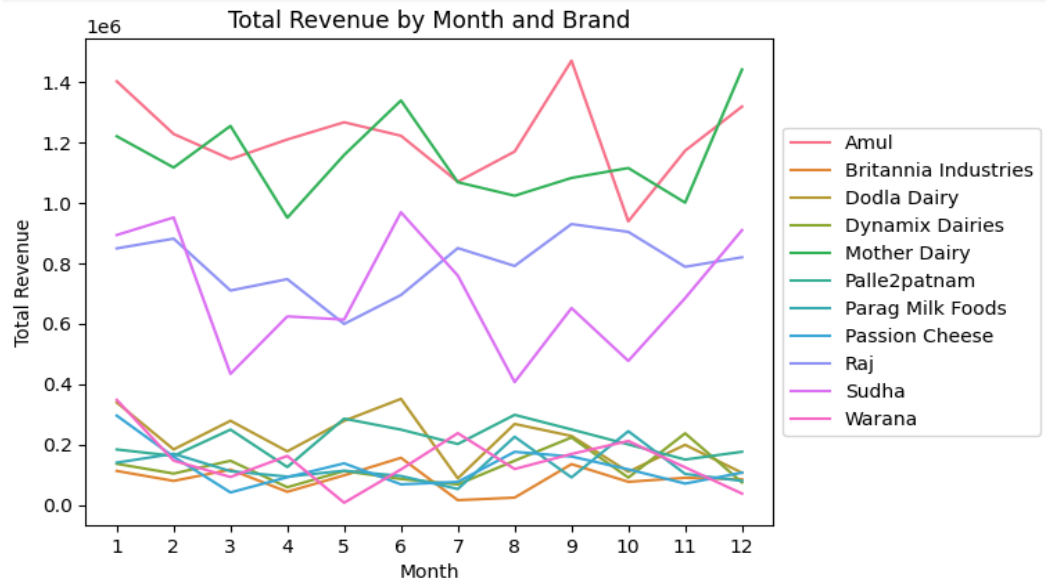


LINE PLOT

```
sns.lineplot(data=df, x='Date', y='Total Value')
plt.title('Line Plot of Total Value over Time')
plt.xticks(rotation=45)
plt.show()
```




```
df['Month']=df['Date'].dt.month
monthly_brand_total=df.groupby(['Month','Brand'])['Total
Revenue'].sum().reset_index()
sns.lineplot(data=monthly_brand_total, x= 'Month', y='Total Revenue',
hue='Brand')
plt.title('Total Revenue by Month and Brand')
plt.xticks(range(1, 13))
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```



```
for obj in obj1:
    for num in num1:
        print(f"{obj} and {num} Analysis")
        print("")
        result = df[[obj, num]].groupby(obj).agg(['mean', 'median', 'min',
'max', 'std', 'count'])
        print(result)
        print("")
        print("\n\n")
```

Location and Total Land Area Analysis

Location	Total Land Area					
	mean	median	min	max	std	count
Bihar	505.064816	503.330	11.86	998.12	290.296662	245
Chandigarh	505.773911	516.520	10.32	999.53	286.849968	519
Delhi	509.527638	531.950	13.87	999.39	290.771416	525
Gujarat	493.443333	490.670	12.64	996.04	272.027588	267
Haryana	481.903439	465.340	13.78	994.13	290.887586	253
Jharkhand	496.482823	499.625	10.91	994.75	285.930636	248
Karnataka	497.195249	443.030	15.16	999.18	292.452830	261
Kerala	516.498313	534.630	14.97	985.12	279.075638	249
Madhya Pradesh	524.648185	553.210	11.32	997.26	290.016260	259
Maharashtra	487.731451	489.950	17.99	995.38	292.120882	255
Rajasthan	484.853672	482.700	10.25	993.92	274.854787	256
Tamil Nadu	539.920269	581.690	11.48	998.55	294.735135	223
Telangana	493.878548	494.000	10.17	992.69	278.629996	248
Uttar Pradesh	521.866920	533.455	13.25	998.53	290.237292	276
West Bengal	486.940705	475.440	11.73	990.93	274.169808	241

Location	count
Bihar	245
Chandigarh	519
Delhi	525
Gujarat	267
Haryana	253
Jharkhand	248
Karnataka	261
Kerala	249
Madhya Pradesh	259
Maharashtra	255
Rajasthan	256
Tamil Nadu	223
Telangana	248
Uttar Pradesh	276
West Bengal	241

Location and Shelf Life Analysis

Location	Shelf Life					
	mean	median	min	max	std	count
Bihar	28.514286	17.0	1	150	32.144602	245
Chandigarh	28.335260	22.0	1	149	28.589886	519
Delhi	29.960000	22.0	1	149	31.771133	525
Gujarat	30.509363	22.0	1	144	30.477943	267
Haryana	25.588933	18.0	1	131	25.077281	253
Jharkhand	29.108871	23.0	1	145	30.266432	248
Karnataka	30.544061	22.0	1	150	32.938915	261
Kerala	28.184739	22.0	1	147	26.818799	249
Madhya Pradesh	29.571429	23.0	1	147	30.516829	259
Maharashtra	29.121569	22.0	1	150	30.049777	255
Rajasthan	27.546875	18.0	1	143	28.546861	256
Tamil Nadu	31.905830	24.0	1	150	33.399373	223
Telangana	26.572581	16.0	1	149	28.657456	248
Uttar Pradesh	30.630435	23.0	1	150	30.962635	276
West Bengal	30.834025	23.0	1	148	33.143838	241

Location and Number of Cows Analysis

Location	Number of Cows					
	mean	median	min	max	std	count
Bihar	53.934694	53.0	10	100	25.631288	245
Chandigarh	53.709056	53.0	10	100	25.213722	519
Delhi	55.352381	56.0	10	100	26.004030	525
Gujarat	52.711610	51.0	10	99	26.389420	267
Haryana	53.213439	53.0	10	100	25.503538	253
Jharkhand	52.943548	51.0	10	100	26.911552	248
Karnataka	55.613027	56.0	10	100	26.382826	261
Kerala	55.389558	56.0	10	100	25.610546	249
Madhya Pradesh	56.536680	56.0	10	99	25.541758	259
Maharashtra	56.827451	57.0	10	100	26.371394	255
Rajasthan	56.859375	56.5	10	100	27.203441	256
Tamil Nadu	57.677130	61.0	10	100	26.762448	223
Telangana	56.068548	57.0	10	100	27.009358	248
Uttar Pradesh	54.884058	54.0	10	100	26.769609	276
West Bengal	53.900415	53.0	11	100	25.288964	241

Location and Product ID Analysis

Location	Product ID					
	mean	median	min	max	std	count
Bihar	5.559184	6.0	1	10	2.884663	245
Chandigarh	5.510597	6.0	1	10	2.782215	519
Delhi	5.480000	5.0	1	10	2.858975	525
Gujarat	5.730337	6.0	1	10	2.819497	267
Haryana	5.600791	6.0	1	10	2.810609	253
Jharkhand	5.274194	5.0	1	10	2.919538	248
Karnataka	5.681992	6.0	1	10	2.929036	261
Kerala	5.534137	6.0	1	10	2.862409	249
Madhya Pradesh	5.177606	5.0	1	10	2.881261	259
Maharashtra	5.388235	6.0	1	10	2.748346	255
Rajasthan	5.691406	6.0	1	10	2.761515	256
Tamil Nadu	5.376682	5.0	1	10	2.843008	223
Telangana	5.737903	6.0	1	10	2.806848	248
Uttar Pradesh	5.329710	5.0	1	10	2.908918	276
West Bengal	5.597510	6.0	1	10	2.867896	241

Location and Quantity Analysis

Location	Quantity					
	mean	median	min	max	std	count
Bihar	485.031918	470.140	4.39	991.29	289.313376	245
Chandigarh	502.013757	516.820	1.17	999.16	285.931201	519
Delhi	503.079581	490.560	1.24	999.80	299.559101	525
Gujarat	501.400524	497.550	5.23	993.85	277.622958	267
Haryana	472.246285	472.750	3.51	991.49	274.498537	253
Jharkhand	496.550565	493.495	7.52	999.78	276.793975	248
Karnataka	501.067778	498.740	1.22	999.87	296.500632	261
Kerala	481.545181	437.060	2.72	999.82	293.361137	249
Madhya Pradesh	522.066293	528.390	4.82	997.79	286.755315	259
Maharashtra	507.645686	526.620	3.10	995.82	293.551817	255
Rajasthan	489.231602	495.615	4.92	992.99	290.510328	256
Tamil Nadu	529.906951	567.880	8.92	987.59	296.775256	223
Telangana	515.309516	518.635	3.46	999.93	287.031776	248
Uttar Pradesh	503.367609	501.845	1.58	997.98	279.420413	276
West Bengal	497.278797	492.870	4.38	998.55	302.088558	241

Location and Price per Unit Analysis

Location	Price per Unit mean	median	min	max	std	count
Bihar	55.093918	53.620	10.81	99.99	25.887363	245
Chandigarh	55.590405	55.770	10.03	99.96	25.974016	519
Delhi	54.650095	54.770	10.16	99.73	25.945825	525
Gujarat	53.254607	50.870	10.71	99.25	26.372699	267
Haryana	53.956364	53.890	10.68	99.70	25.790044	253
Jharkhand	55.453589	55.845	10.33	99.49	25.630910	248
Karnataka	54.422299	54.510	10.32	99.78	25.326854	261
Kerala	54.148916	52.590	10.05	99.52	26.773213	249
Madhya Pradesh	54.120541	53.180	11.20	99.71	26.461480	259
Maharashtra	55.471608	58.630	10.76	99.51	25.431041	255
Rajasthan	51.086133	49.435	11.14	99.30	25.082309	256
Tamil Nadu	56.824215	57.910	10.41	99.65	26.354563	223
Telangana	53.563427	52.840	10.11	99.78	27.001263	248
Uttar Pradesh	57.497283	58.300	10.16	99.96	25.850110	276
West Bengal	56.155228	56.350	10.19	99.94	26.313209	241

Brand and Price per Unit Sold Analysis

Brand	Price per Unit Sold mean	median	min	max	std
Amul	54.435508	53.770	6.12	103.89	25.965367
Britannia Industries	51.334318	49.235	8.29	100.49	27.492805
Dodla Dairy	53.860135	51.735	7.37	101.56	26.078537
Dynamix Dairies	56.108774	59.250	10.72	101.22	26.476048
Mother Dairy	54.388366	54.430	5.94	102.98	25.905275
Palle2patnam	52.302227	50.540	7.62	102.33	26.987878
Parag Milk Foods	59.732549	65.380	8.83	100.43	27.572955
Passion Cheese	55.292292	56.665	8.48	100.66	25.901043
Raj	55.543693	54.140	5.21	102.42	25.895629
Sudha	55.066019	53.550	5.61	104.51	26.576286
Warana	58.445000	57.945	8.27	104.15	26.802298

Location and Total Value Analysis

Location	Total Value mean	median	min	max	std
Bihar	26833.525081	22780.67000	97.5096	96137.3400	21733.714240
Chandigarh	27801.181202	21679.86680	42.5165	99036.3696	21384.104613
Delhi	27002.981423	22583.19420	70.6056	91387.7055	21368.522973
Gujarat	26224.816396	21453.43440	141.5270	94083.7590	19745.619601
Haryana	25143.051233	19315.00100	141.2775	86051.2128	20197.645099
Jharkhand	27358.513414	23067.79680	134.3072	90520.2535	20072.593156
Karnataka	27237.041443	21954.65740	46.0428	89527.1400	22431.732821
Kerala	26906.944076	20270.74500	93.6700	94594.4951	22938.361223
Madhya Pradesh	28460.424510	22378.34640	298.0910	97631.1921	22712.455448
Maharashtra	27278.727043	22587.61230	47.4300	93567.2472	21095.511446
Rajasthan	25549.524894	19522.68900	106.3200	96528.5579	21766.852362
Tamil Nadu	30288.612331	22409.19180	548.7581	85997.4090	23740.247269
Telangana	27265.725519	21888.96160	120.0274	91933.5264	21034.128517
Uttar Pradesh	29112.023851	25893.61935	73.2962	94246.0402	21727.140237
West Bengal	28079.172329	21865.62200	218.8841	90044.1984	22868.960448

Brand and Total Revenue Analysis

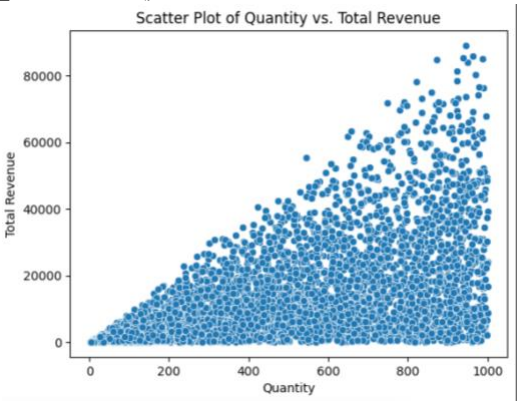
Brand	Total Revenue mean	median	min	max	std
Amul	13873.565489	8187.660	16.30	84838.46	14688.162512
Britannia Industries	11699.114205	6904.600	222.09	55540.50	12352.926512
Dodla Dairy	11733.591757	7109.240	26.77	60313.95	13476.468411
Dynamix Dairies	13961.177170	8632.020	12.54	66323.46	15479.728595
Mother Dairy	13630.481653	8879.460	14.99	85352.94	14549.264974
Palle2patnam	11985.730427	6655.000	19.66	78214.50	14488.714338
Parag Milk Foods	14869.585392	11131.690	23.89	62856.66	14034.184572
Passion Cheese	15610.970312	12095.525	34.22	73718.57	15100.366592
Raj	13957.339620	8376.000	21.12	89108.90	15192.797462
Sudha	12916.835741	7495.485	17.46	81350.52	14193.208583
Warana	17014.476442	11922.365	256.86	76247.22	16331.988186

A higher standard deviation suggests greater fluctuations in sales or prices, while a lower value indicates consistency.The heighest value recorded for the metric in each group reveals the peak sales performance for a product or category.

Highlights the typical sales performance for different dairy products or storage conditions.Mean value represents the average value of the numerical metric (e.g., sales quantity or sales price) for each group.

SCATTER PLOT

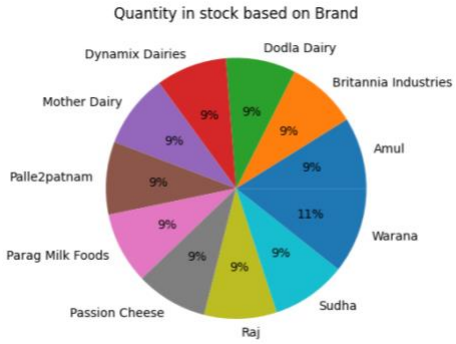
sns.scatterplot(data=df, x='Quantity', y='Total Revenue')
plt.title('Scatter Plot of Quantity vs. Total Revenue')
plt.show()

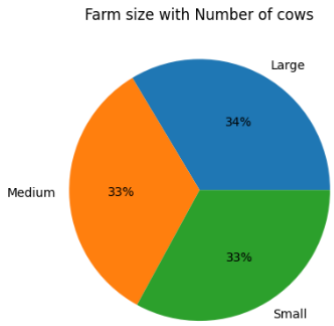


sns.scatterplot(data=df, x='Quantity Sold', y='Price per Unit Sold')
plt.title('Scatter Plot of Quantity Sold vs. Price per unit sold')
plt.show()

PIE PLOT

plt.pie(df.groupby('Brand')['Quantity in Stock'].mean(),labels=df.groupby('Brand')['Quantity in Stock'].mean().index,autopct='%1.0f%%')
plt.title('Quantity in stock based on Brand')
plt.show()





The pie chart analysis revealed that the brand Warana holds the highest quantity in stock, significantly standing out among all brands. In contrast, the stock quantities for the other brands are relatively similar and evenly distributed.

There are 34% Cows present in the farm where the size is large and other medium and small size farms contains only 33%.From this analysis through pie plot it is clear that the farm that has the count of cows higher have highest production than rest of the farms.

MULTIVARIATE ANALYSIS

result = df.groupby(['Product Name', 'Brand'])['Quantity'].mean()
result

Product Name	Brand	Quantity
Butter	Amul	510.040648
	Mother Dairy	488.143419
	Parag Milk Foods	509.980098
	Warana	589.990385
Buttermilk	Amul	472.735393
	Mother Dairy	480.897627
	Raj	509.273750
	Sudha	478.874828
Cheese	Amul	539.968378
	Britannia Industries	482.887955
	Dynamix Dairies	497.864057
	Passion Cheese	532.974687
Curd	Amul	501.201157
	Mother Dairy	550.146891
	Raj	529.345798
	Sudha	542.251417
Ghee	Amul	481.784632
	Mother Dairy	489.248269
	Raj	488.087019
	Sudha	496.518586
Ice Cream	Amul	519.270472
	Dodla Dairy	469.156000
	Mother Dairy	482.970893
	Palle2patnam	502.157500
Lassi	Amul	537.340769
	Mother Dairy	501.790280
	Raj	510.878413
	Sudha	468.110825
Milk	Amul	481.683131
	Mother Dairy	546.530171
	Raj	513.448142
	Sudha	455.975900
Paneer	Amul	482.702737
	Mother Dairy	461.727563
	Raj	482.402973
	Sudha	484.769655
Yogurt	Amul	442.467500
	Dodla Dairy	462.085128
	Mother Dairy	543.680309
	Palle2patnam	487.301441

The analysis indicates that the product *Butter* is produced by four different brands. Among them, Warana stands out as the brand with the highest production volume.

The analysis shows that Buttermil*k* is predominantly produced by the brand *Raj*, while other brands contribute significantly less to its production.

The analysis reveals that Cheese is predominantly produced by the brand Amul, with Passion Cheese following closely behind in production volume.

Mother Dairy is the brand which produces curd more. Ghee is highly produced by the brand Sudha.

The production of Ice creams is mostly popular for the brand Amul. The production of milk which is highly in demand is produced by the brand Mother Diary where both the curd and milk is highly produced.

result = df.groupby(['Product Name', 'Brand'])['Quantity Sold'].mean()
result

Product Name	Brand	
Butter	Amul	233.027778
	Mother Dairy	232.094017
	Parag Milk Foods	262.372549
	Warana	293.211538
Buttermilk	Amul	248.359551
	Mother Dairy	256.000000
	Raj	233.205357
	Sudha	189.405172
Cheese	Amul	260.810811
	Britannia Industries	239.170455
	Dynamix Dairies	252.566038
	Passion Cheese	284.333333
Curd	Amul	238.834711
	Mother Dairy	269.151261
	Raj	235.487395
	Sudha	262.200000
Ghee	Amul	237.926316
	Mother Dairy	251.615385
	Raj	243.413462
	Sudha	246.737374
Ice Cream	Amul	266.839623
	Dodla Dairy	214.190476
	Mother Dairy	252.053571
	Palle2patnam	257.750000
Lassi	Amul	274.239316
	Mother Dairy	233.728972
	Raj	267.793651
	Sudha	225.711340
Milk	Amul	265.020202
	Mother Dairy	258.726496
	Raj	279.725664
	Sudha	217.350000
Paneer	Amul	263.410526
	Mother Dairy	214.680672
	Raj	269.675676
	Sudha	236.431034
Yogurt	Amul	224.062500
	Dodla Dairy	232.282051
	Mother Dairy	279.670103
	Palle2patnam	228.324324

result=df.groupby(['Product Name','Storage Condition'])['Quantity Sold'].mean().sort_values(ascending=False)
result

Product Name	Storage Condition	
Milk	Tetra Pack	263.534314
Cheese	Frozen	260.747253
Butter	Frozen	260.703540
Cheese	Refrigerated	258.488584
Yogurt	Frozen	256.573529
Lassi	Refrigerated	252.194631
Curd	Refrigerated	251.388309
Milk	Polythene Packet	249.293333
Ice Cream	Frozen	247.706856
Butter	Refrigerated	247.117073
Ghee	Ambient	245.057214
Paneer	Refrigerated	244.741497
Buttermilk	Refrigerated	230.809195
Yogurt	Refrigerated	224.905579

dtype: float64

result=df.groupby(['Product Name','Sales Channel'])['Price per Unit Sold'].mean().sort_values(ascending=False)
result

		Price per Unit Sold
Product Name	Sales Channel	
Butter	Wholesale	59.397063
Milk	Retail	57.862555
Butter	Retail	57.862075
Paneer	Wholesale	57.563987
Lassi	Online	57.386748
Curd	Retail	57.258674
Buttermilk	Online	56.877467
Butter	Online	56.559922
Ghee	Wholesale	56.445166
	Online	55.755726
Buttermilk	Retail	55.740432
Yogurt	Retail	55.427434
Ghee	Retail	55.210448
Buttermilk	Wholesale	55.030000
Lassi	Retail	54.603869
Cheese	Wholesale	54.164218
Milk	Wholesale	54.016483
Cheese	Online	54.011692
Yogurt	Online	53.638492
Paneer	Online	53.468403
Curd	Wholesale	53.448288
Milk	Online	53.432245
Ice Cream	Retail	52.939110
Paneer	Retail	52.882189
Ice Cream	Wholesale	52.834820
Curd	Online	52.757105
Yogurt	Wholesale	52.183711
Lassi	Wholesale	51.870748
Cheese	Retail	51.843952
Ice Cream	Online	49.547174

result=df.groupby(['Customer Location','Product Name','Price per Unit'])['Reorder Quantity'].mean().sort_values(ascending=False).head(10)
result

Customer Location	Product Name	Price per Unit	Reorder Quantity
Bihar	Curd	76.25	199.95
Telangana	Ghee	21.89	199.92
Haryana	Milk	29.93	199.81
Gujarat	Buttermilk	85.61	199.77
Jharkhand	Ghee	23.64	199.71
West Bengal	Lassi	29.07	199.71
Tamil Nadu	Ghee	68.66	199.68
Jharkhand	Buttermilk	47.50	199.59
Chandigarh	Paneer	27.30	199.47
Telangana	Buttermilk	56.35	199.45

dtype: float64

result=df.groupby(['Product Name','Total Value','Price per Unit'])['Reorder Quantity'].mean().sort_values(ascending=False).head(10)
result

Product Name	Total Value	Price per Unit	Reorder Quantity
Curd	14770.3875	76.25	199.95
Ghee	10052.5447	21.89	199.92
Milk	26422.2040	29.93	199.81
Buttermilk	43975.2887	85.61	199.77
Lassi	18023.1093	29.07	199.71
Ghee	17482.0164	23.64	199.71
	65932.8248	68.66	199.68
Buttermilk	7071.3250	47.50	199.59
Paneer	7100.1840	27.30	199.47
Buttermilk	11263.8015	56.35	199.45

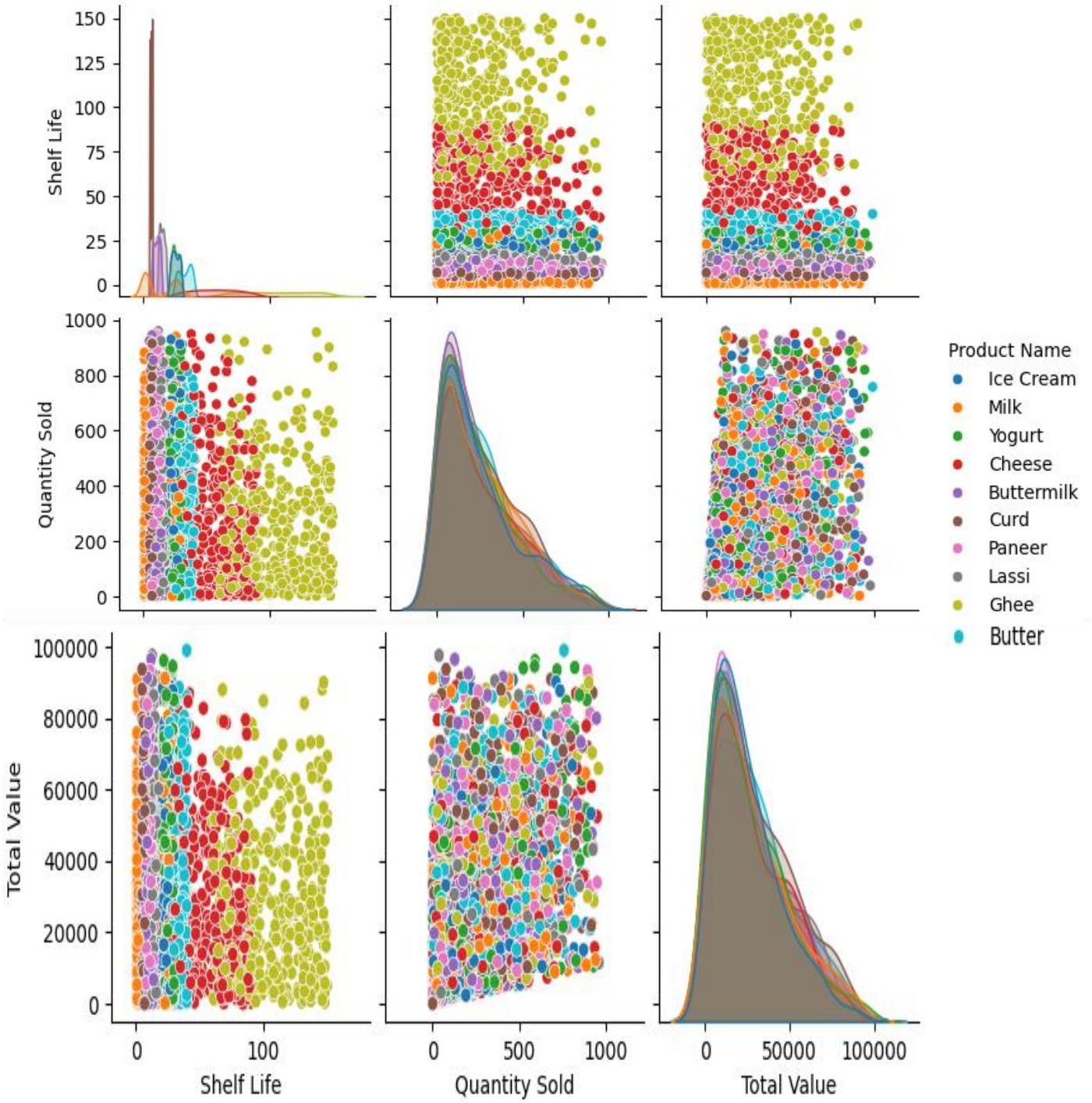
dtype: float64

result=df.groupby(['Product Name','Shelf Life','Quantity Sold'])['Total Value'].mean().sort_values(ascending=False).head(10)
result

Product Name	Shelf Life	Quantity Sold	Total Value
Butter	40	758	99036.3696
Buttermilk	12	129	96528.5579
Yogurt	22	592	96137.3400
	29	589	94594.4951
	27	820	94246.0402
	29	524	94083.7590
Curd	5	79	93714.6730
Yogurt	29	718	93567.2472
Paneer	11	893	93254.2650
Buttermilk	13	207	92625.3576

dtype: float64

PAIR PLOT


Shelf Life
Quantity Sold
Total Value
Product Name
Ice Cream
Milk
Yogurt
Cheese
Buttermilk
Curd
Paneer
Lassi
Ghee
Butter
0
100
500
1000
50000
100000
20000
40000
60000
80000
200
400
600
800
25
50
75
125
150

CONCLUSION

- The majority of dairy farms are concentrated in Delhi, which also accounts for the highest sales. This suggests a strong correlation between farm location and customer demand.
- Tamil Nadu has the largest farm areas and the highest number of cows, indicating significant production capacity.
- Curd is the most produced product, showcasing its popularity and high demand among consumers.
- Paneer and cheese are the least produced products, reflecting lower consumer demand or production focus.
- Amul emerges as the most popular brand with the highest product count, showcasing its dominance in the market. Britannia Industries has the least product count, indicating limited market presence.
- The most preferred storage condition is a refrigerator, highlighting the perishable nature of dairy products.
- Ghee has the highest shelf life, making it a sustainable product for long-term storage, while curd has the least shelf life, requiring quicker turnover.
- August sees the highest sales, likely driven by seasonal demand, while April records the lowest sales.
- The average sales volume remains moderate, with most sales counts falling within the range of 5000 units.
- Milk has the highest reorder quantity, indicating its status as a staple product with consistent demand. Ice cream have the least reorder quantities, possibly reflecting their niche market appeal.
- Most customers who are highly interested in dairy products are located in Delhi, aligning with the production and sales trends in the region.
- Increase production of high-demand products like curd and milk to cater to market needs.
- Explore opportunities to boost the demand for underperforming products like cheese and paneer
- Strengthen partnerships with popular brands like Amul while exploring strategies to enhance the presence of underrepresented brands
- Focus on optimizing refrigeration facilities to support the storage of perishable products like curd and milk.
- Develop targeted marketing campaigns for August to capitalize on peak sales and create strategies to boost sales during low-demand months like April.
- Focus on premium products like flavored milk, probiotic yogurt, and organic dairy products to attract health-conscious consumers.
- Customize product offerings to suit regional preferences, such as paneer for northern markets or curd for southern markets.
- Increase the number of retail outlets in underserved regions to tap into new customer bases.
- Offer incentives to farms that produce high-quality milk to maintain product standards.