FRESHMART

Maximising Total Sales Revenue through Smarter Retail Decisions

FRESHMART ANALYSIS:

1.INTRODUCTION:

FreshMart is a fast-growing grocery retail chain based in the United States, serving thousands of customers across various cities and countries. Known for its wide product range and affordable pricing, FreshMart has built a strong presence in both urban and suburban markets

As the company prepares for its next phase of growth, leadership wants to focus not just on adding new stores, but on increasing Total Sales Revenue from its existing network. This means a better understanding what drives revenue, from which products perform well, to how different regions, customer segments, and sales staff contribute to the bottom line.

Here our main objective is to increase the Total Sales Revenue through different types of way available for us.The company also believes there are many unseen opportunities to increase the Sales revenue generated.

2. OBJECTIVES:

Our main goal is to analyze the monthly sales of Freshmart to uncover actionable insights that can help increase the Total Sales Revenue

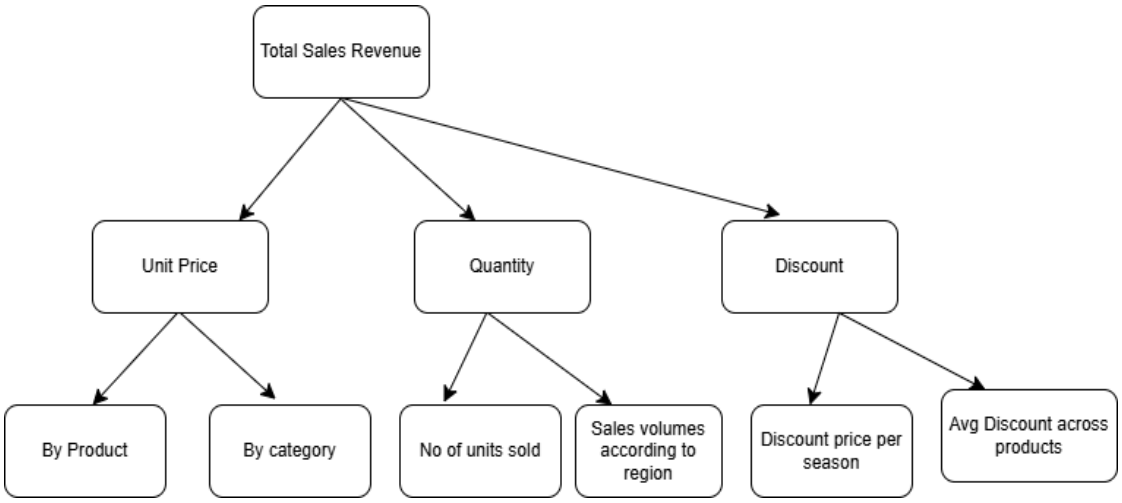
To uncover actionable insights from the FreshMart dataset that will help maximize Total Sales Revenue by analyzing:

- Product performance across time
- Customer purchase behavior
- Employee contributions
- City/country-wise revenue
- Discount effectiveness
- Employee performance
- Product categorization

KEY METRIC:- Total Sales Revenue= Unit Price × Quantity × (1 - Discount)

This metric helps us measure how much money Freshmart is making from selling products after applying discounts.

METRIC TREE:



This tree visually decomposes the primary revenue metric into its core components. Each branch can be optimized independently to grow revenue. The metric tree is used to identify which sub-metric price, volume, or discount is influencing revenue positively or negatively.

Business Impact

Improving Total Sales Revenue- even by small percent–can generate hundreds of millions in incremental profit annually for a brand like FreshMart.
The FreshMart data analysis provides a direct and measurable impact on the company's strategic revenue.
By breaking down Total Sales Revenue into its influencing components–price,quantity and discount–this project helps uncover where sales performance can be improved without increasing costs.

Key insights like identifying high performing products,cities and employees enable the company to allocate resources more efficiently. Optimizing discount strategies ensures that margin erosion is minimized while sales remain competitive. Furthermore, understanding customer behavior and segmenting buyers based on their revenue contribution allows for more personalized marketing and retention strategies.

Identifying which product types and categories generate the highest revenue will allow FreshMart to focus on high value items, streamline production, and allocate resources more effectively, ensuring maximum returns on investment.

Implementing the findings from this analysis will:

- Improve inventory planning for high-revenue categories
- Reduce unnecessary discounts while preserving volume
- Boost employee performance through targeted training and incentives
- Increase profitability from existing store networks
- Strengthen FreshMart's data-driven culture for long-term decision making

Dataset Overview:

- **Dataset Name :** FRESHMART DATASET
- **Number of Rows :** 6,857,636
- **Number of Columns :** 32
- **Description:**This dataset captures key metrics related to retail transactions and operational performance for FreshMart, a grocery retail chain. Each row in the dataset represents an individual transaction enriched with related attributes such as product details, pricing, customer information, and discount levels.The data is organized into seven structured ones, each focusing on a core entity within FreshMart's operations:

Each table is related to other as given below:

Table 1	Key Column(s)	Table 2	Key Column(s)
sales	CustomerID	customers	CustomerID
sales	ProductID	products	ProductID
sales	SalesPersonID	employees	EmployeeID
sales	SalesDate	discounts	StartDate, EndDate
employees	CityID	cities	CityID
customers	CityID	cities	CityID
cities	CountryID	countries	CountryID
products	CategoryID	categories	CategoryID

Column Definitions:

The below column names are as per my ones which i have given them:

1.sales_freshmart.csv – Records each sale with quantity, price, discount, and total revenue.

- **SalesID**: Unique identifier for each transaction.
- **SalesPersonID**: Refers to the employee responsible for the sale (linked to employees.csv).
- **CustomerID**: The customer who made the purchase (linked to customers.csv).
- **ProductID**: Product that was sold (linked to products.csv).
- **Quantity**: Number of units purchased in the transaction.
- **Discount**: Discount applied to the product during this sale, in decimal format
- **TotalPrice**: Final transaction value after discount.
- **SalesDate**: Timestamp of the transaction.
- **TransactionNumber**: Unique transaction reference ID.

2.products_freshmart.csv – Contains product-level details like name, price, category, and classification.

- **ProductID**: Unique identifier for each product.
- **ProductName**: Name of the product
- **Price**: Unit price of the product before any discounts.
- **CategoryID**: Reference to the product category (linked to categories.csv).
- **Class**: Product classification such as Standard or Premium.
- **ModifyDate**: Date when product information was last updated.
- **Resistant**: Resistance properties
- **IsAllergic**: Indicates whether the product contains allergens.
- **VitalityDays**: Shelf life or freshness period of the product in days

3.customers_freshmart.csv – Includes customer demographics and location references.

- **CustomerID**: Unique identifier assigned to each customer.
- **FirstName**: The customer's first name.
- **MiddleInitial**: Middle initial of the customer (may be missing in some cases).
- **LastName**: The customer's last name.
- **CityID**: Reference to the city where the customer resides (linked to cities.csv).
- **Address**: Full residential address of the customer.

4.employees_freshmart.csv – Provides information on sales staff responsible for transactions.

- **EmployeeID**: Unique identifier for each employee.

- **FirstName**: First name of the employee.
- **MiddleInitial**: Middle initial (may not be consistently filled).
- **LastName**: Last name of the employee.
- **BirthDate**: Employee's date of birth.
- **Gender**: Gender of the employee
- **CityID**: Refers to the city the employee is assigned to (linked to cities.csv).
- **HireDate**: Date the employee was hired by FreshMart

5.categories_freshmart.csv – Defines each product's category for aggregation and analysis.

- **CategoryID**: Unique identifier for each product category.
- **CategoryName**: Name of the product category (e.g., Dairy, Beverages).

6.cities_freshmart.csv – Lists city names with corresponding population and region codes.

- **CityID**: Unique identifier for each city.
- **CityName**: Name of the city
- **Zip Code**: Represents the city's approximate population.
- **CountryID**: Foreign key referencing the country (linked to countries.csv).

7.countries_freshmart.csv – Captures country-level data to support regional comparisons.

- **CountryID**: Unique identifier for each country.
- **CountryName**: Full name of the country (e.g., United States).
- **CountryCode**: Two-letter ISO-style country code (e.g., US).

Steps:

Data Cleaning and Preparation:

Data cleaning and preparation is one of the crucial steps in carrying out the FreshMart analysis.

It is carried out on the above 7 datasets of FreshMart analysis.

Each column of the dataset was examined first to check the datatypes ,kind of values they hold including the duplicates,missing and null values also.They are also corrected by performing by some functions.

Step 1: Importing Data and Required Python Libraries

Data Import: The datasets were sourced from a shared Google Drive link and imported into Google Colab for analysis.

Python Libraries Used: Libraries such as Pandas, NumPy, Matplotlib, and Seaborn were imported to perform data cleaning, handle null/missing values, process merges, and generate visual insights.

Step 2: Dataset Overview & Initial Exploration

Shape: All datasets were checked for shape using .shape. For example, sales dataset has 6758125 rows.

Data Types: Data types of each column were verified using .info() and corrected as needed. For example, SalesDate was converted to datetime, and ID fields were cast as str.

Missing Values: Missing values were counted using .isnull().sum() for each dataset.

- SalesDate in the sales dataset had 67526 nulls.
- MiddleInitial in the customers dataset had 977 missing entries.
- CountryCode had 1 missing value.

Duplicates: Duplicates were checked using .duplicated().sum(). Minor duplicates found in sales and customers were removed.

Step 3: Handling Missing Values

- **Numerical Columns**: No missing values were found in key numerical columns such as Price, Quantity, or TotalPrice.
- **Categorical Columns**:
- MiddleInitial was left blank.
- SalesDate nulls were filled as one of the dates as it is essential for time-based analysis.

Step 4: Modifying & Standardizing Data Values

Some columns needed value normalization and formatting:

- **Date Fields**: SalesDate, HireDate, and BirthDate were converted to proper datetime objects.
- **Datasets Merging**: Proper joins were performed using foreign keys like CityID, CategoryID, and EmployeeID to ensure each sale is linked to a valid product, customer, city, and employee.

Step 5: Datatype Correction

- **Verify Data Types**: Data type of each attribute was reviewed and if wrong converted to suitable data type after outlier and null value handling.
- **Converting Data Types**: All the identifier columns were assigned string datatype, all the date related columns were assigned datetime datatype, numerical columns were assigned integer datatype and categorical were converted to string type.

Step 6: Feature Engineering

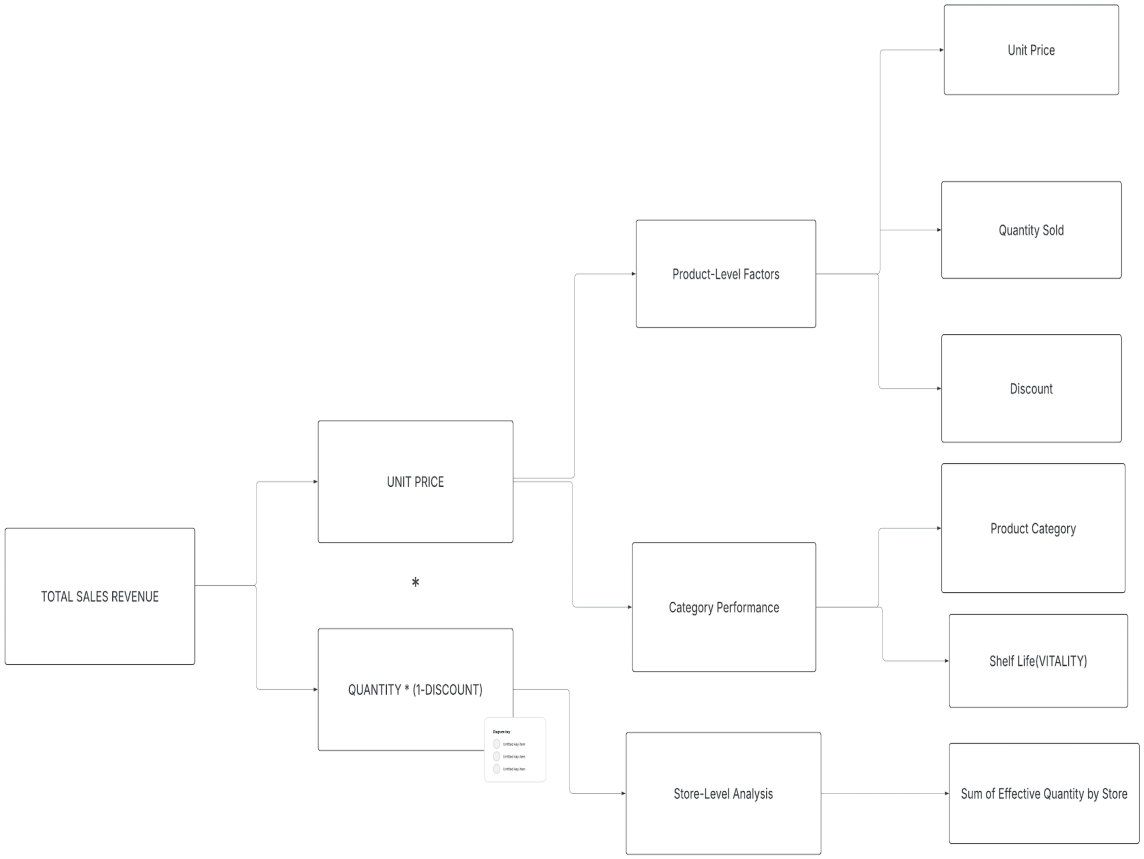
These are additional datasets derived via merging or transformation:

- **sales_with_products** – Merged sales_freshmart with product and category data.
- **Sales_per_revenue** – Sales data enriched with product revenue calculations.
- **Sales_per_customer** – Combines sales with customer information.
- **Sales_per_employee** – Merges employee data to track revenue per salesperson.
- **Sales_per_city** – Adds city-level revenue tracking to customer-sales merged data.
- **product_category** – Merged product data with category info.
- **product_category_revenue** – Revenue aggregated by product and category.
- **category_revenue** – Summarized total revenue per product category.
- **avg_discount** – Averages discounts.

Metric Tree:

Total Sales Revenue – FreshMart(Metric Tree)

This metric tree breaks down the key components contributing to Total Sales Revenue, showing both core formula components and influencing factors.



Exploratory Data Analysis (EDA):

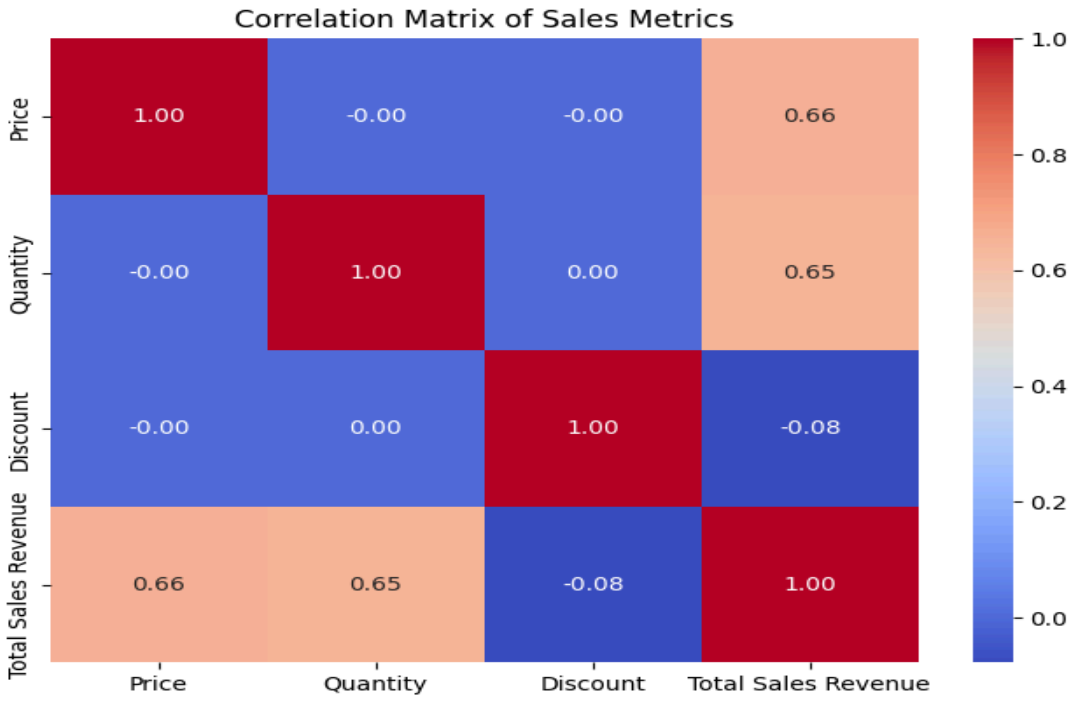
As we know the Exploratory Data Analysis phase is focused mainly on uncovering meaningful patterns,trends and relationships across FreshMart's data.

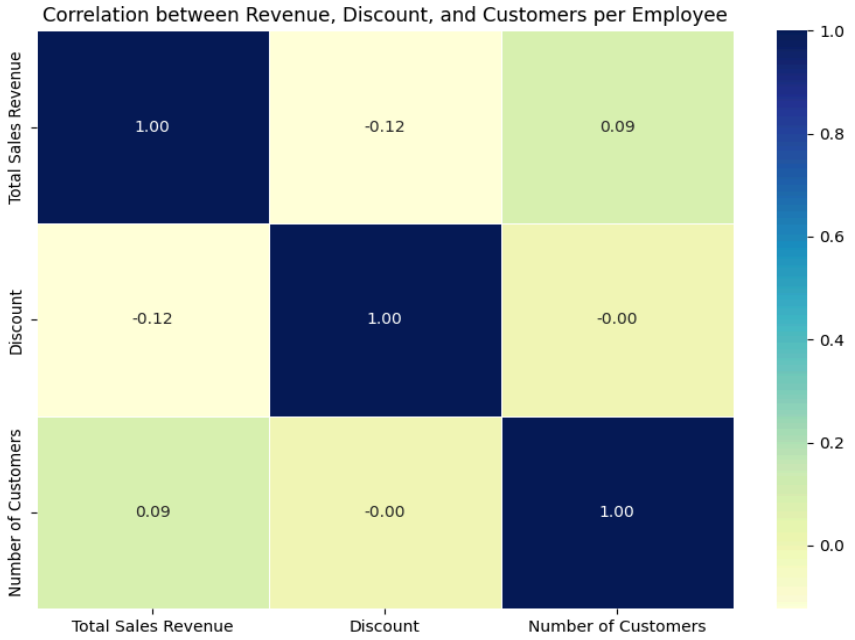
Key variables can include Price,Quantity,Discount and Total Sales Revenue which were analyzed by some the visualization techniques.

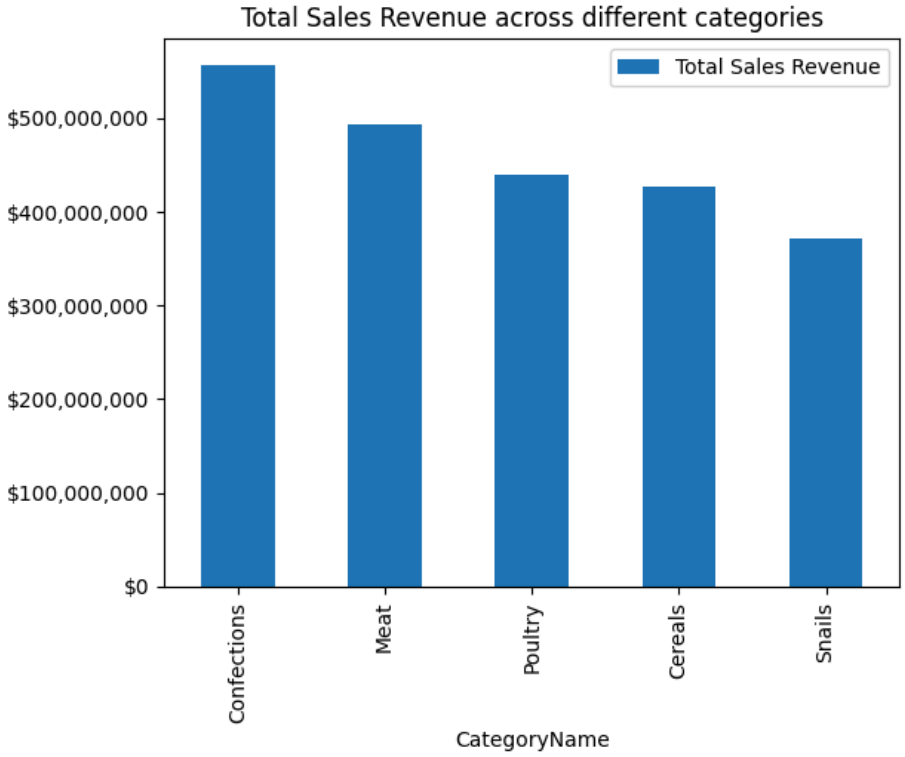
The EDA is overall laid the foundation for hypothesis testing and KPI tracking by highlighting the key revenue driving segments within FreshMart datasets.

Key Summary Statistics:

METRIC	VALUE(NO)
TOTAL REVENUE	4332445646.06
TOTAL QUANTITY	87882708
AVERAGE DISCOUNT	0.03
NO OF UNIQUE PRODUCTS	452
NO OF CATEGORIES	66
HIGHEST REVENUE CATEGORY	Confections
NO OF EMPLOYEES	276
NO OF UNIQUE CUSTOMERS	98759




Correlation between Revenue, Discount, and Customers per Employee
Total Sales Revenue
Discount
Number of Customers
1.00
-0.12
0.09
-0.12
1.00
-0.00
0.09
-0.00
1.00
Total Sales Revenue
Discount
Number of Customers
1.0
0.8
0.6
0.4
0.2
0.0


Total Sales Revenue across different categories
Total Sales Revenue
$500,000,000
$400,000,000
$300,000,000
$200,000,000
$100,000,000
$0
Confections
Meat
Poultry
Cereals
Snails
CategoryName

From the above observations we can say some of the important key summary statistics and from correlation matrix heatmap also:

- The correlation matrix says that the higher the Quantity the higher the total sales revenue would be.
- It also says that the more the discount the less is total sales revenue
- We can also say that Confections is the highest revenue producer in categories.
- We can also see that the high priced products are not purchased frequently.

Hypothesis Formulation and Testing:

1)Product Category wise performance:

Hypothesis 1: Some Category wise products contribute more revenue than the others.

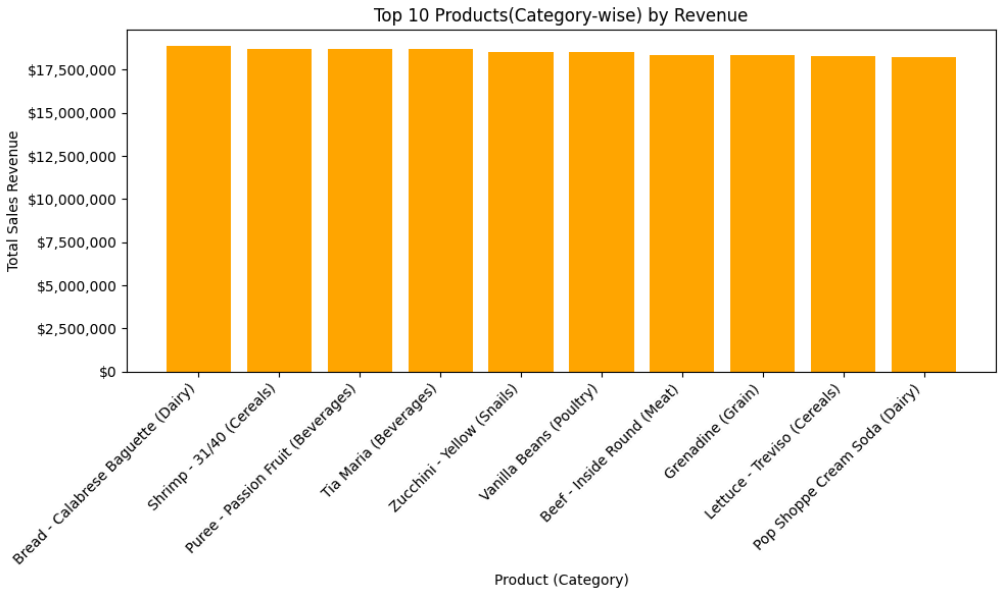
	ProductName	CategoryName	Total Sales Revenue
56	Bread - Calabrese Baguette	Dairy	18868838
332	Shrimp - 31/40	Cereals	18721942
299	Puree - Passion Fruit	Beverages	18703480
379	Tia Maria	Beverages	18685124
451	Zucchini - Yellow	Snails	18551657
393	Vanilla Beans	Poultry	18530394
29	Beef - Inside Round	Meat	18334978
172	Grenadine	Grain	18331168
211	Lettuce - Treviso	Cereals	18321893
287	Pop Shoppe Cream Soda	Dairy	18241372

Observations:

- From the above observation we are showing the top 10 products(category-wise) which produce the highest total sales revenue.
- The highest top contributor as we can see is Bread which is from a Diary Category which has the highest Sales Revenue among all the products(category-wise)

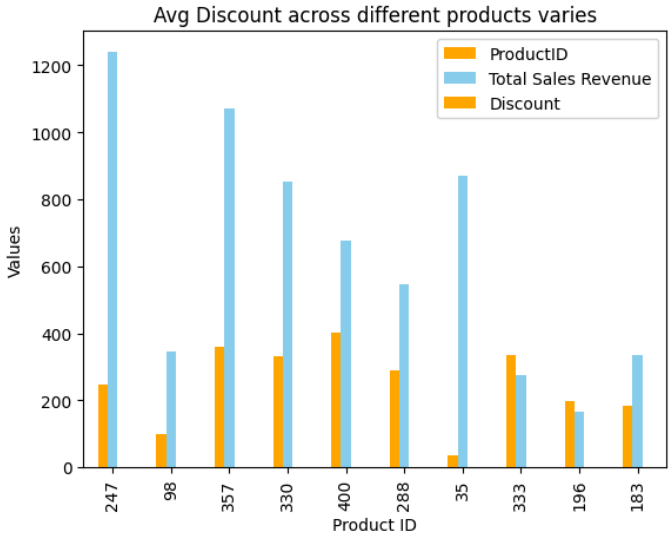
Recommendations:

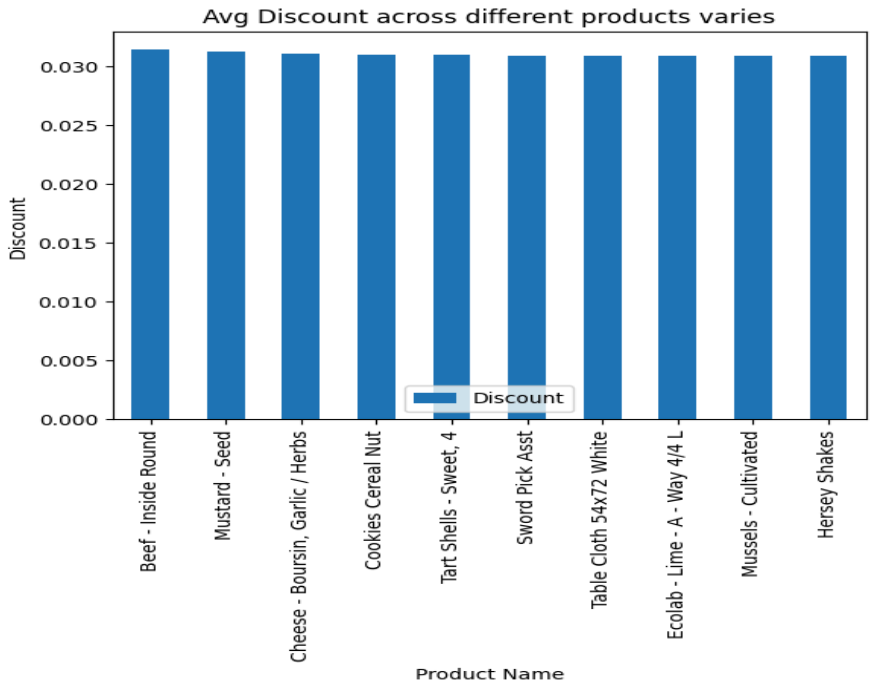
- We can say that we should highly concentrate on producing much more quantities of bread to help increase the total sales revenue.
- We should also produce more Shrimp in quantities as we can see they are nearly competing with the bread in sales revenue which can massively increase the total sales revenue



2)Discount across different products:

Hypothesis 2: The average discount varies across different products which can affect the total sales revenue





Observations:

- As we can see from the above bar graphs the average discount varies for different products.
- It also gives a massive impact on the total sales revenue for different products as the more the average discount the less the sales revenue for that product will be there.
- I can also say that products with minimum average discount will be the highest sales revenue contributors.
- We should observe which products have average discounts but also are producing more sales revenue and produce them more in quantity to increase the sales revenue.

Recommendations:

- Focus on the products that achieve high sales revenue with minimal discounts, as they offer better profit than the others
- Avoid applying high discounts on products where it leads to a total downfall in the total sales revenue of that product
- Also see the products with much discount but their sales revenue is also high when compared to others and promote those .

3)Does more quantity mean more revenue?

Hypothesis 3: More quantity of products means more revenue.

Observations:

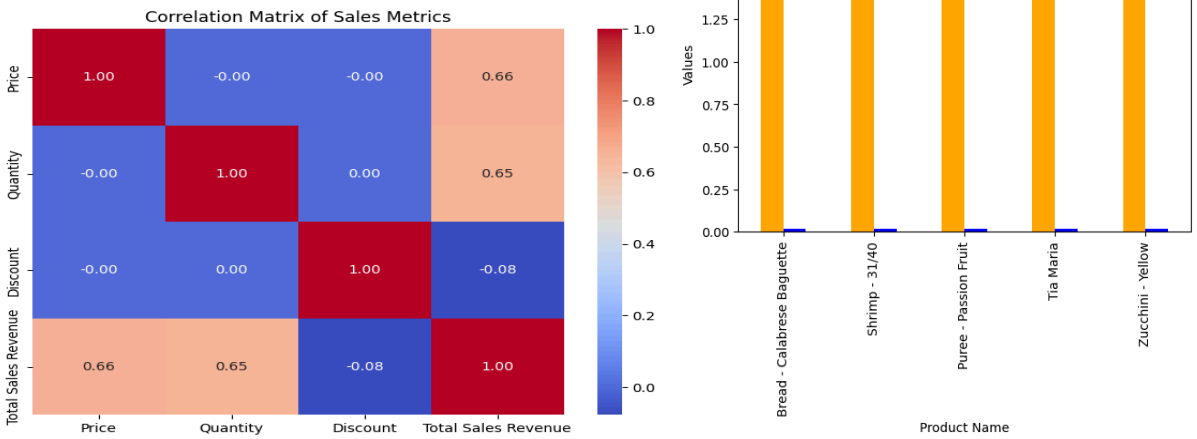
- Products with highest quantity production will produce more revenue when compared to products with lower quantity .
- We can see their positive correlation from the above plots.
- By this we can prove the hypothesis that more quantity means more revenue generated.

For example we got bread as the highest revenue generator in the above one and as you can see here the quantity of bread produced also is more.

Recommendations:

- Focus on scaling the production of high-demand products like bread to maximize revenue generation.

- Continuously monitor the quantity-to-revenue ratio to ensure that increased production leads to proportional revenue growth



4)Total Sales Revenue varies across different cities:

Hypothesis 4:Different cities have different sales revenue

Observations:

- Here we are providing a pie chart that gives us information on the top 5 cities across Sales revenue in percentage.
- This percentage mainly depends on the products sold in those cities ,the number of customers in those cities and also the employee performance.
- These are all the important factors that help in analyzing the sales revenue across top 5 cities .

Recommendations:

- We have to observe what are the factors that are helping these cities generate the highest sales revenue than the other ones.
- By this we can also use those factors on the other cities which help in gaining more sales revenue
- We have to also see the store performance and compare them with other ones to know which helps in increasing the sales revenue percentage.